Publishing while female

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

Erin Hengel†

August 2018

I use readability scores to test if women's writing is held to higher standards 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 almost two times higher in published articles than in their pre-print drafts; (iii) women's writing gradually improves but men's does not, so the readability gap grows over authors' careers. Using a subjective expected utility framework, I show that tougher editorial standards and/or biased referee assignment are uniquely consistent with authors' observed choices. A conservative estimate derived from the model suggests higher standards cause senior female economists to write at least 7 percent more clearly than they otherwise would. I also document evidence that higher standards affect behaviour and lower productivity. First, female-authored papers take six months longer in peer review. Second, women appear to gradually adapt to higher standards *in* peer review by writing more readably *before* it. The latter response disguises external thresholds as personal choice; the former reduces women's output. Both whitewash discrimination. More generally, tougher standards impose a quantity/quality tradeoff that helps explain academia's "Publishing Paradox". Since evidence of this tradeoff is present beyond academia, higher standards may also contribute to women's lower productivity in many occupations.

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

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

1 Introduction

Ladies, our papers aren't published that often in "top-four" economics journals.[1] In 2015, the average share of female authors per paper was 15 percent. Only eight percent were majority female-authored.[2] Just four percent were written entirely by women. The *Quarterly Journal of Economics* did not publish a single exclusively female-authored paper between 2015–2017 (inclusive). In several recent years, *Econometrica* and the *Journal of Political Economy* have not either.

These statistics are uncomfortable, but their causes are myriad: lower publishing rates, career choices, motherhood and, possibly, bias. In lab experiments women are subject to tougher standards. Their qualifications and ability are underestimated (Foschi, 1996; Grunspan et al., 2016; Moss-Racusin et al., 2012; Reuben et al., 2014). Female-authored manuscripts are evaluated more critically (Goldberg, 1968; Krawczyk and Smyk, 2016; Paludi and Bauer, 1983); when collaborating with men, women are given less credit (Heilman and Haynes, 2005; Sarsons, 2017).

This paper uses five measures of writing clarity to suggest 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. I use a dynamic model of an author's decision-making process to show that tougher editorial standards and/or biased referee assignment are the only explanations consistent with men's and women's diverging choices. A conservative estimate derived from the model suggests higher standards in peer review cause senior female economists to write at least seven percent more clearly than they otherwise would.

I also document evidence that higher standards confound productivity measurement and their own identification. First, higher standards presumably delay review. To test this hypothesis, I estimate the gender difference in submit-accept times at *Econometrica*, controlling for, among other things, motherhood and childbirth. Female-authored papers spend six months longer in peer review.

Second, discrimination is often observed in—or on the path toward—an equilibrium state. My own data suggest women gradually adapt to higher standards *in* peer review by writing more readably *before* it. This implies that women adjust to biased treatment in ways that partially—or even totally—confuse it with voluntary choice. Studies that analyse only one slice of the equilibrium path must take into account how it would have evolved had discrimination not occurred. Otherwise, they risk underestimating it, misallocating responsibility or even concluding bias against men.

Higher standards impose a quantity/quality tradeoff that plausibly contributes to academia's "Publishing Paradox" and "Leaky Pipeline".[3] Spending more time revising old research means there's less time for new research. Fewer papers results in fewer promotions, possibly driving women into fairer fields. Since there is evidence of this tradeoff in a variety of occupations—*e.g.*, doctors, real estate agents and airline pilots—higher standards could distort women's productivity,

[1] Economics journals in the "top-four" are: *American Economic Review*, *Econometrica*, *Journal of Political Economy* and *Quarterly Journal of Economics*.

[2] For comparison, 28, 26 and 13 percent of assistant, associate and full professors, respectively, are women (Lundberg, 2017). "Majority female-authored" refers to papers with a ratio of female authros strictly above 50 percent.

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

more generally.

Prior research typically concludes that there is no evidence of gender bias in academic peer review (see, *e.g.*, Blank, 1991; Borsuk et al., 2009; Gilbert et al., 1994). Yet studies on the topic tend to analyse a single indicator (acceptance rates) from a narrow context (publication outcomes).[4] I ask a different question. Men's and women's papers may be published at comparable rates, but do referees consistently apply identical standards in how they scrutinise and evaluate their work? 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 more revisions and have less patience deciphering their 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).[5]

To test whether journals hold female- and male-authored papers to identical writing standards, 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 9,122 article abstracts published in the *American Economic Review* (*AER*), *Econometrica* (*ECA*), *Journal of Political Economy* (*JPE*) and *Quarterly Journal of Economics* (*QJE*).

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; nor (iv) generated by factors correlated with gender but really related to knowledge, intelligence and creativity.

Additionally, the gender readability gap substantially widens *during* peer review. To establish this result, I compare National Bureau of Economic Research (NBER) working papers to their final, published versions.[6] 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 do not partially correlate with the latter's timing, a widening gap suggests a causal link.[7]

Two explanations could account for these findings: either women voluntarily write better papers—*e.g.*, because they're more sensitive to referee criticism—or better written papers are women's response to external circumstances they do not control. Both imply women spend too

[4] Of course, gender neutral acceptance rates do not imply gender neutral outcomes if women's papers are better than men's.

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

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

[7] Although the causal link I establish at this step is only with peer review—and *not* with editor and/or referee bias—I also find no evidence of a gender readability gap in the (small) sample of NBER working paper–published article pairs subjected to double-blind review pre-internet (Section 4.3.3). This result suggests editorial bias does indeed play a role.

much time rewriting old papers and not enough time writing new papers.[8] The appropriate policy response, however, depends on the cause. To distinguish between them, I model an author's decision-making process within a subjective expected utility framework.

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

Theorem 1 establishes sufficient conditions to demonstrate double standards are present in academic peer review: (1) experienced women write better than equivalent men; (2) women improve their writing over time; (3) female-authored papers are accepted no more often than equivalent male-authored papers. Estimates from pooled subsamples at fixed publication counts suggest (1) and (2) hold. On average, women's writing gradually gets better but men's does not; between authors' first and third published articles, the readability gap increases by up to 12 percent. Although my data do not identify probability of acceptance, conclusions from extensive study elsewhere are clear: "there are no sex differences in acceptance rates." (Ceci et al., 2014, p. 111).[9]

To interpret the relationship as causal, however, technically requires that each of Theorem 1's conditions 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. To account for this, I match prolific female authors to similarly productive male authors on characteristics that predict the topic, novelty, and quality of research.

Conditions (1) and (2) were satisfied in 65 percent of matched pairs. In three-quarters of those, the member discriminated against was female. A conservative estimate derived from the model suggests higher standards cause senior female economists to write at least seven percent more clearly than they otherwise would.[10]

As a final exercise, I show suggestive evidence that higher standards affect women's behaviour and their productivity measurement. First, writing and revising a paper is hard work; making sentences even marginally more readable takes time. Thus, higher standards imply female-authored papers spend longer under review. I test this hypothesis using submit-accept times from *Econometrica*. Indeed, female-authored papers spend ***six months longer*** in peer review. The effect persists across a range of specifications and, in addition to other factors, controls for motherhood, childbirth, citations and field.

Second, to tease out the direct effects of higher standards—readability changes made in peer review—from its indirect effects—readability changes made before peer review—I compare papers pre- and post-review over increasing publication counts. In authors' earliest papers, the

[8] Alternatively, men spend too much time writing new papers and not enough time rewriting old papers.

[9] See also Section 4.4.2 for references to other research supporting this claim.

[10] This estimate averages results over all five scores and assumes women are accepted in a subset of states in which men are accepted and within pair differences are zero for the 30–40 percent of matched pairs that fail to satisfy Conditions (1) and (2). Alternative estimates based on weaker assumptions are shown in Table 8 and Appendix M.4, Table M.7. While seven 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.

readability gap exclusively emerges *during* peer review (direct effect). In fact, there is no significant gender difference between draft readabilities in men's and women's first top publications. In later papers, however, women write well upfront; the gap chiefly materialises *before* peer review (indirect effect). The latter observation implies an equilibrium state in which constrained outcomes are indistinguishable from voluntary choice. The former indicates women do not initially anticipate higher standards in peer review—meaning papers by junior female economists could be (desk) rejected at higher rates.[11]

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). The present paper, in contrast, exclusively explores discrimination's impact on the behaviour and choices of people discriminated against.

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. As I demonstrate here, however, a traditional "outcome test" combined with repeatedly observing authors choices also reveals bias by editors and/or referees.

Outcome tests provide powerful evidence of discriminatory treatment—but only if they credibly isolate group differences in the observed equilibrium from those that would have occurred in the non-discriminatory counterfactual equilibrium.[12] Because it is generally assumed that men and women possess different preferences, knowledge about underlying processes and/or personality traits exacerbated by imperfect information, assuming no gender difference in the counterfactual equilibrium is difficult to justify.

Instead, I use information revealed by authors' repeated readability choices to reconstruct it: assuming preferences are fixed over time, earlier choices provide an upper bound on the impact intrinsic preferences play in gender readability gaps; assuming authors update beliefs about the relationship between readability and acceptance rates means later choices are made with accurate beliefs. The basic logic, moreover, applies equally well to any situation where people are repeatedly judged on and respond to feedback about some quantifiable component of their output.

The second advantage of analysing discrimination from the perspective of people discriminated against is that it forces us to think more deeply about its impact on, *inter alia*, occupational choice, worker motivation, human capital investment and, especially, productivity measurement.[13] This paper joins a small, emerging empirical literature examining these effects (*e.g.*, Glover et al., 2017; Lavy and Sand, 2015; Parsons et al., 2011).[14]

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.[15] Fewer papers justifies fewer promotions.[16] If women seek fairer employment elsewhere—

[11] This conclusion follows directly from Theorem 1. Section 4.5 provides limited evidence and further discussion.

[12] For a more detailed discussion of outcome tests, see Ayres (2001, Ch. 9).

[13] This was a major theme in earlier theoretical work. Lundberg (1991) and Lundberg and Startz (1983) show that discrimination can lead to suboptimal human capital investment when firms more reliably assess the productivity of members in one group relative to another. Occupational segregation by gender is one potential outcome in Goldin (2014b)'s "pollution" theory of discrimination—*e.g.*, male economists resist the entry of female economists (because their presence "pollutes" the profession's prestige), so women sort (or are pushed) into "feminised" research fields.

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

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

[16] Evidence on whether female academics are hired and promoted at lower rates is mixed. One study suggests so-

or quit the labour force entirely—it feeds a "Leaky Pipeline".[17]

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)[18] 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 (Funk and Perrone, 2017; 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 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);[19] patients treated by female physicians are less likely to die or be readmitted to hospital (Tsugawa et al., 2017); female pilots are involved in fewer fatal accidents (Bazargan and Guzhva, 2011; Vail and Ekman, 1986);[20] 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, 2017; Cortés and Pan, 2016; Pertold-Gebicka et al., 2016); women's tendency to under negotiate pay (Babcock and Laschever, 2003)[21] 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

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).

[17] It may also manifest itself in more indirect ways, *e.g.*, by impacting the types of research women conduct, the effort they put into their work or the amount and extent of service to the profession they choose to undertake.

[18] 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).

[19] 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).

[20] 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.

[21] A more recent study suggests women do ask for higher pay—they just don't get it (Artz et al., 2018).

"femininity": female physicians consult longer with patients (Roter and Hall, 2004); female politicians fundraise more intensely (Jenkins, 2007);[22] 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).[23]

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, 2014a; Goldin and Katz, 2016), preferences (for a review, see, *e.g.*, Blau and Kahn, 2017) and policy design (Antecol et al., 2018). 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 in every single one.

Furthermore, my results advocate using caution when employing 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 market 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.[24]

This paper makes two final contributions. First, it adds to extensive (ongoing) research into peer review and academia's "Publishing Paradox". 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 editorial patterns (Card and DellaVigna, 2013; Clain and Leppel, 2018; Ellison, 2002), bias in editorial decisions (Abrevaya and Hamermesh, 2012; Bransch and Kvasnicka, 2017; Card and DellaVigna, 2017) and female economists' lagging productivity and general underrepresentation (Bayer and Rouse, 2016; Ductor et al., 2018; Ginther and Kahn, 2004).

Second, my findings emphasise the importance of transparency and monitoring. The least intrusive antidote to implicit bias is simple awareness and constant supervision. Both factors foster accountability and encourage neutrality (Foschi, 1996). Monitoring referee reports is difficult, but

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

[23] 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).

[24] 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).

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

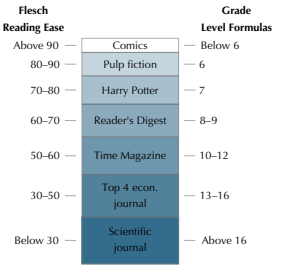

FIGURE 1: Calculating and interpreting readability scores

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

it isn't impossible—especially if peer review were open. As discussed in Section 5, 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 discusses readability scores. Section 3 describes the data and the gender representation of articles published in top economics journals. Analyses and results are presented in Section 4. I close with a summary, discussion (Section 5) and conclusions (Section 6).

2 Readability scores

Advanced vocabulary and complicated sentences are two strong predictors of text difficulty (Chall and Dale, 1995). Hundreds of formulas exploit this relationship to measure so-called "readability". I concentrate on the most widely used, tested and reliable formulas for adult reading material: Flesch Reading Ease, Flesch-Kincaid, Gunning Fog, SMOG (Simple Measure of Gobbledegook) and Dale-Chall (DuBay, 2004).[25] Each are listed in Figure 1.

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

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

Readability scores correlate with (i) oral reading fluency,[27] (ii) human judgement, (iii) reading comprehension tests and (iv) the cloze procedure.[28] The dark blue box plots in Figure 2 summarise

[25] A sixth commonly used measure is the Lexile Framework. Because its formula and software are proprietary, I do not include it in the analysis.

[26] Included in this sample are between-score correlations found in two non-published studies—the present paper (correlations range from 0.53 to 0.97) and Benoit et al. (2017).

[27] Oral reading fluency is generally measured as the number of words read aloud correctly per minute.

[28] The cloze procedure ranks passages of text according to average readers' ability to correctly guess randomly deleted words.

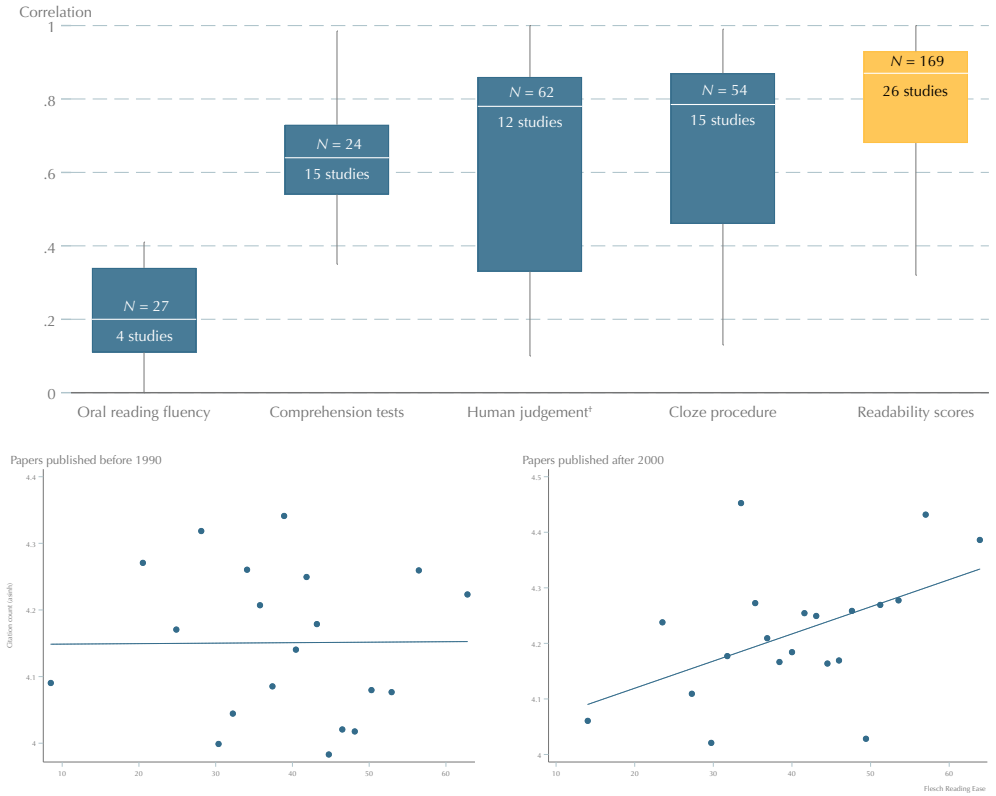

FIGURE 2: Readability score validity

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

167 correlations in 38 published cross-validation studies.

Other studies have validated readability scores against surrogate measures of reading comprehension. More readable high school and college-level correspondence courses have higher completion rates (Klare and Smart, 1973). More readable academic journals enjoy larger readerships (Richardson, 1977; Swanson, 1948); their most readable articles win more awards (Sawyer et al., 2008), are downloaded more often (Guerini et al., 2012)[29] and cited more frequently (see Figure 2).[30]

Thanks to high predictive power and ease of use, readability formulas are widely employed in education, business and government. The U.S. Securities and Exchange Commission encourages clearer financial disclosure forms benchmarked against the Gunning Fog, Flesch-Kincaid and Flesch Reading Ease scores (Cox, 2007). The formulas have also guided readability assess-

[29] In a blog post, Lukas Püttmann compares abstract readability to page views of VoxEU.org columns: more readable columns are viewed three percent more often (Püttmann, 2017).

[30] Evidence from other studies linking readability and citations is weaker. Lei and Yan (2016) find a positive yet non-significant relationship between readability and citations in information journals. In finance journals, Berninger et al. (2017) shows citations positively correlate with abstract readability but negatively correlate with readability in the body of a paper. Laband and Taylor (1992) did not find any relationship between the two. In my own data, there is a positive relationship in papers published after 1990—and particularly those published post–2000—but no relationship before that (Figure 2).

ments of, *inter alia*, standardised test questions (Chall et al., 1983; Chall et al., 1977), medical inserts (*e.g.*, Wallace et al., 2008), technical manuals (*e.g.*, Hussin et al., 2012; Klare and Smart, 1973), health pamphlets (*e.g.*, Foster and Rhoney, 2002; Meade and Byrd, 1989) and data security policies (Alkhurayyif and Weir, 2017).

In research, readability scores are considered objective proxies for "complexity". Enke (2018) controls for language sophistication using the Flesch Reading Ease formula in a study of moral values in U.S. presidential elections. Spirling (2016) employs the same score to show that British parliamentarians simplified speeches to appeal to less educated voters in the in the wake of the Great Reform Act. Legal research has found that judges are more reliant on legislative history when interpreting complex legal statutes, as measured by the Flesch-Kincaid formula (Law and Zaring, 2010).[31] In finance, the scores have linked clarity of financial communication to better firm and market financial health (Biddle et al., 2009; Jansen, 2011; Li, 2008), larger investment and trading volume (De Franco et al., 2015; Lawrence, 2013; Miller, 2010; Thörnqvist, 2015) and lower demand for—albeit higher reliability of—outside research by sell-side analysts (Lehavy et al., 2011).[32]

2.1 Measurement error

Readability scores fail to capture many elements relevant to reading comprehension, including grammar—*e.g.*, active vs. passive tense (Coleman, 1964; Coleman, 1965)—legibility—*e.g.*, typeface or layout—and content—*e.g.*, coherence, organisation and general appeal (Armbruster, 1984; Kemper, 1983; Kintsch and Miller, 1984; Meyer, 1982). Nevertheless, "long sentences generally correspond to complex syntactic structures, infrequent words generally refer to complex concepts, and hard texts will generally lead to harder questions about their content" (Kintsch and Miller, 1984, p. 222).[33]

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

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

Conditional on accurate calculation, readability scores combine very precise estimates of vocabulary complexity with almost perfect measures of sentence length (for a discussion, see Chall and Dale, 1995).[34] The weighted average of these two variables is informative in much the same way that inferences about readability are. Thus, measurement error related to (c) should only shift superficial interpretation of observed gender differences—from "women are better writers" to "women use simpler words and write shorter sentences"[35]—but leave conclusions deduced from them intact.

[31] Long and Christensen (2011) investigate whether a legal brief's readability score correlates with its success on appeal (it does not).

[32] See Loughran and Mcdonald (2016) for a review of finance and accounting research using readability measures.

[33] Combining readability scores with measures that capture these features does not, however, increase their predictive power (see, *e.g.*, Kemper, 1983).

[34] The relationship may even be causal: words *become* shorter when used more frequently (Zipf, 1935). For example, "television" was shortened to "TV" and "telly" in American and British English, respectively (example from Chall and Dale, 1995).

[35] Or even "women write shorter sentences and use words with fewer syllables, fewer polysyllabic word and more words on the Dale-Chall list of easy words".

Nevertheless, I try to minimise measurement error from (c) by using abstracts as textual input. Abstracts are self-contained, universally summarise the research and are the first and most frequently read part of an article (King et al., 2006).[36] Moreover, their layout is relatively standardised compared to other parts of a manuscript—abstracts are generally surrounded by ample whitespace and most editorial management systems anyway reproduce them in pre-formatted cover pages. These factors suggest a relatively homogenous degree of review across journals and subject matter and limit the impact that physical layout, figures and surrounding text have on readability.

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

Second, I carefully proofread the text in order to identify (and fix) remaining transcription errors,[37] eliminate non-sentence-ending full stops,[38] and replace typesetting code—typically used to render equations[39]—with equivalent unicode characters.[40] Readability scores were determined using the modified text.

Finally, some programs that calculate scores rely on unclear, inconsistent and possibly inaccurate algorithms to count words and syllables, identify sentence terminations and check whether a word is on Dale-Chall's easy word list (for a discussion, see Sirico, 2007). To transparently handle these issues and eliminate ambiguity in how the scores were calculated, I wrote the Python module `Textatistic`. Its code and documentation are available on GitHub; a brief description is provided in Appendix B. For added robustness, I re-calculate scores and replicate most results using the `R readability` package (Appendix I). Coefficients are very similar to—and (to my chagrin) standard errors universally smaller than—those presented in the body of the paper.

3 Data

The data include every English article published in *AER*, *Econometrica*, *JPE* and *QJE* between January 1950 and December 2015 (inclusive). The largest sample is from *Econometrica* which consistently published abstracts with its articles prior to 1950. *JPE* added them in the 1960s and *QJE* in 1980. *AER* came last in 1986.[41] Table 1 displays data coverage by journal and decade. Figure 3 breaks down the sample's abstract readability by publication year and primary *JEL* classification.

The analysis in Section 4.3 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

[36] Prior research has also found that authors write in a stylistically consistent manner across the abstract, introduction and discussion sections of a paper (Hartley et al., 2003b; Plavén-Sigray et al., 2017).

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

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

[39] Equations in abstracts generally only occur in *Econometrica* articles published before 1980.

[40] 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.

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

Table 1: Article count, by journal and decade

Decade	*AER*	*ECA*	*JPE*	*QJE*	Total
1950–59		120			120
1960–69		343	184		527
1970–79		660	633	1	1,294
1980–89	180	648	562	401	1,791
1990–99	476	443	478	409	1,806
2000–09	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.

eventually matched to 1,988 NBER working papers—approximately one-fifth of the data.[42] Bibliographic information and abstract text were scraped from www.nber.org. Descriptive statistics are provided in Section 4.3.1.

The analysis in Section 4.6 compiles submit-accept times at *Econometrica*—the only journal of the four to make disaggregated data on its revision process publicly available.[43] I extracted this information from digitised articles using the open source command utility `pdftotext`. Section 4.6 displays and discusses basic summary statistics.

Other variables used in the analysis include, *inter alia*, editor fixed effects, institution fixed effects,[44] author productivity fixed effects, proxies for English fluency, citation counts,[45] and controls for motherhood and childbirth (Section 4.6, only). See Appendix C for further information on how each was calculated.

3.1 Gender

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

Determining the "gender" of a paper is not nearly as straightforward. For solo-authored manuscripts—of which there are 4,016 in the sample—gender corresponds to the sex of the

[42] 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.

[43] 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.

[44] Institution controls are coded dynamically (see Appendix C for details on how they are calculated). Earlier versions of this paper (available on my website) report results using static controls.

[45] Unless otherwise noted, I control for citations using the inverse hyperbolic sine of citation count (asinh). The December 2017 version of this paper (available on my website) uses raw counts.

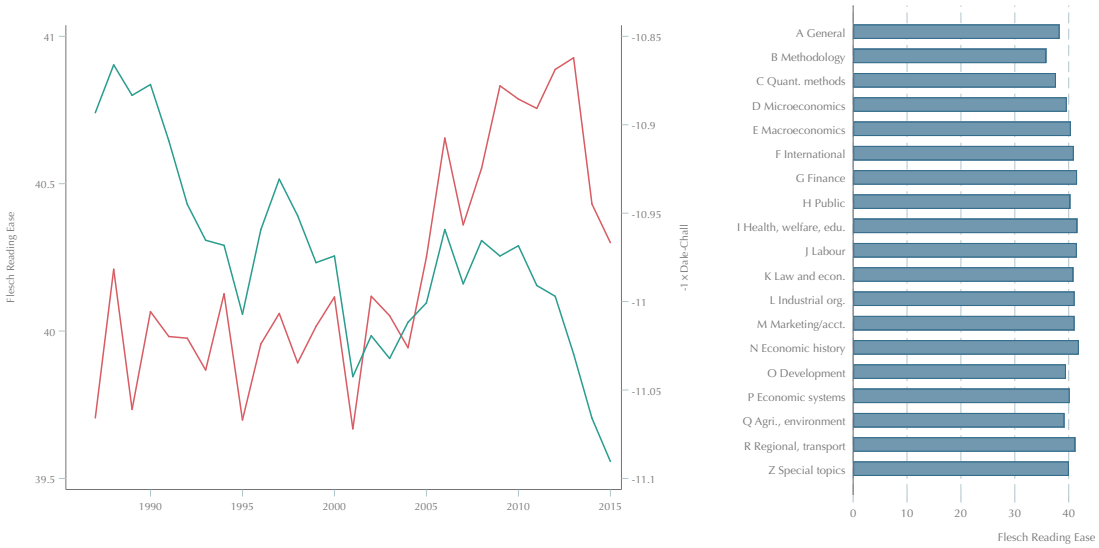

FIGURE 3: Readability by year and *JEL* code

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

author. Unfortunately, top economics journals have collectively published just 266 by women. Only a slightly larger number were written entirely—*or even mostly*—by women (Figure 4).[46] Proportions are similar when the sample is restricted to later years: *QJE* did not publish a single exclusively female-authored paper between 2015–2017 (inclusive); in eight of the last fifteen years covered by the data (2001–2015), *Econometrica*, *JPE* or both did not either.

A greater number of papers (1,172) are authored by at least one woman. To take advantage of the information contained in this larger sample, Blank (1991) classified all such papers as "female". I opt instead for a less inclusive and continuous measure of gender: the proportion of female authors.[47]

This approach assumes a linear relationship between a paper's readability and its gender composition.[48] For robustness, I repeat most analyses (a) on the subset of papers authored by a single gender; (b) using a binary variable equal to one if at least one author is female; and (c) using a binary variable equal to one if at least half of all authors are female. Standard errors from (a) tend to be larger; those from (b) and (c) usually smaller. In general, however, results do not meaningfully change (Appendix J).

4 Analyses and results

In Section 4.1 and Section 4.2, I scrutinise readability at the article- and author-level, respectively. The results establish a gap does exist and rule out obvious confounding factors—*e.g.*, women

[46] 312 papers in the sample were authored entirely by women. Women made up more than 50 percent of all authors in another 47. In 35 observations, a woman was the lead author—*i.e.*, the first author was female in a paper with authors listed non-alphabetically or in which contributions were explicitly noted.

[47] A gender readability gap—if it exists—is presumably a function of (i) the probability a passage of text was written and/or revised by a female co-author; and (ii) referees' beliefs about female authors' contributions to the writing and/or revision of a co-authored paper. I assume the intersection of (i) and (ii) is positively related to the ratio of female authors on a paper based on prior research suggesting that co-authors—regardless of seniority—share responsibility for writing and (especially) revising collaborative work (Hart, 2000; Kumar and Ratnavelu, 2016).

[48] The analysis in Section 4.2 suggests the relationship is increasing and convex.

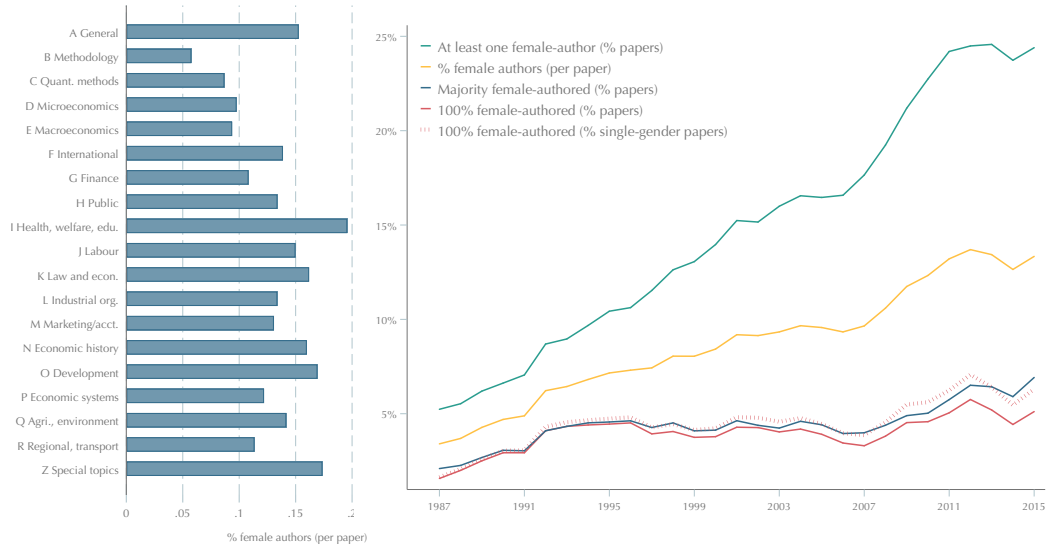

FIGURE 4: The representation of women in top economics journals

Notes. Graphs illustrate the representation of female authors in articles published in a top-four economics journal. Figure on the left is the average share of female authors per paper broken down by primary *JEL* category (5,216 articles); figure on the right displays five year moving averages of various indicators over the period 1987–2015 (6,176 articles).

writing on easier topics, editorial policies in earlier eras, author-specific productivity effects, *etc.*

In Section 4.3, I causally link a portion of the gap to the peer review process. To do so, I match published articles—which have gone through peer review—to earlier, draft versions of the same papers—which have not. On average, the peer review process appears to cause about 40 percent of the gap in readability between men and women.

Section 4.4 is the final step in my identification strategy; it causally links the gap to referees and/or editors. Reasons *why* peer review causes women to write more readably sort into two mutually exclusive groups: (i) factors within their control—*e.g.*, sensitivity to criticism—versus (ii) factors outside their control—*i.e.*, gender bias by referees and/or editors. I develop a dynamic model of an author's decision-making process in order to distinguish between (i) and (ii) (Section 4.4.1). According to my estimate, higher standards from editors and/or referees cause senior female economists to write at least seven percent more clearly than they otherwise would (Section 4.4.2).

I conclude Section 4 by documenting suggestive evidence that discrimination affects behaviour and lowers productivity. In Section 4.5, I investigate how women react as they update beliefs about referees' expectations. in Section 4.6, I test one observable repercussion from subjecting female authors to higher standards—prolonged peer review.

4.1 Article-level analysis

Table 2 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 3 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

Table 2: Textual characteristics per sentence, by gender

	Men	Women	Difference
No. characters	134.72	130.38	4.34***
	(0.43)	(1.46)	(1.57)
No. words	24.16	23.08	1.08***
	(0.08)	(0.27)	(0.29)
No. syllables	40.65	38.68	1.97***
	(0.13)	(0.45)	(0.48)
No. polysyllabic words	4.69	4.31	0.38***
	(0.02)	(0.07)	(0.08)
No. difficult words	9.38	8.92	0.47***
	(0.03)	(0.12)	(0.13)

Notes. Sample 9,122 articles. 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.

accordingly.[49]

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.[50] Percentage-wise, women write about 1–2 percent better than men.[51]

Column (2) includes 63 year dummies; column (3) adds another 182 journal and year interaction dummies; columns (4) and (5) introduce dynamic institution effects, quality controls—citation count (asinh) and 30 max. T_j effects (maximum co-author lifetime publication count for paper j)—and a dummy variable capturing English fluency.[52] Coefficients and standard errors in columns (2)–(5) are fairly similar to those in column (1).

The coefficients on the journal dummies in (2) are presented in Appendix K. 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 (7) and (8) control for primary *JEL* classification. (7) includes 19 fixed effects for primary *JEL* categories; (8) includes 718 effects for tertiary categories. Since only post–1990 *JEL* classifications are used, estimates in both columns exclude over 40 percent of the data. (Due to small sample sizes, (8) includes 561 articles from *AER Papers & Proceedings*.)[53] Coefficients

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

[50] 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.

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

[52] 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 3. In addition to the control from English fluency presented here, see Hengel (2016, pp. 35–36) for further evidence that the female authors in my data are no more or less likely to be native English speakers.

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

TABLE 3: Gender differences in readability, article-level analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Flesch Reading Ease	0.89*	0.86*	0.81	0.87*	1.05**	0.65	0.64	0.86
	(0.49)	(0.48)	(0.50)	(0.50)	(0.51)	(0.59)	(0.58)	(0.72)
Flesch-Kincaid	0.17	0.16	0.16	0.18	0.20	0.21	0.23*	0.23
	(0.11)	(0.11)	(0.11)	(0.11)	(0.12)	(0.14)	(0.13)	(0.15)
Gunning Fog	0.31**	0.30**	0.31**	0.32**	0.35**	0.36**	0.34**	0.33*
	(0.12)	(0.12)	(0.12)	(0.12)	(0.13)	(0.16)	(0.14)	(0.17)
SMOG	0.20**	0.20**	0.20**	0.21**	0.23**	0.21*	0.19*	0.21
	(0.09)	(0.09)	(0.09)	(0.09)	(0.10)	(0.12)	(0.11)	(0.13)
Dale-Chall	0.10**	0.10**	0.09**	0.10**	0.12**	0.12**	0.11*	0.13**
	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)	(0.06)
Editor effects	✓	✓	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓	✓	✓
Institution effects				✓	✓	✓	✓	✓
Quality controls					✓[1]	✓[1]	✓[1]	✓[1]
Native speaker					✓	✓	✓	✓
JEL (primary) effects							✓	
JEL (tertiary) effects								✓

Notes. 9,122 articles in (1)–(5); 5,216 articles in (6) and (7); 5,777 articles—including 561 from *AER Papers & Proceedings* (see Footnote 53)—in (8). Coefficients in (6) are estimated on the same sample as (7) but without *JEL* fixed effects. Figures represent the coefficient on female ratio from an OLS regression on the relevant readability score. Quality controls denoted by ✓[1] include citation count (asinh) and max. T_j fixed effects. Standard errors clustered on editor in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

and standard errors are roughly equivalent to those in column (6), which estimates the gender readability gap using the same restricted sample from (7) but omits field dummies.

Appendix E explores field in more detail. Conditional on other explanatory variables, however, I find little evidence that it drives the gender readability gap observed in Table 3.

4.2 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 × year, institution and English fluency dummies; quality controls—citation count (asinh) 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 $1/N_j$.[54] Standard errors are adjusted for two-way clustering on editor and author.

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

TABLE 4: Gender differences in readability, author-level analysis

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Female ratio (women)	2.54**	0.36*	0.64**	0.47**	0.26**
	(1.02)	(0.20)	(0.25)	(0.19)	(0.10)
Female ratio (men)	0.72	0.11	0.16	0.10	0.11
	(1.37)	(0.26)	(0.31)	(0.21)	(0.10)
Female ratio×male	−1.81	−0.25	−0.48	−0.36	−0.15
	(1.52)	(0.31)	(0.37)	(0.26)	(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	−20.15	−15.88	−16.93	−19.78	−20.96
Order 2	0.57	−0.33	0.10	0.30	−0.50
N_j	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓
Quality controls	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]
Native speaker	✓	✓	✓	✓	✓

Notes. Sample 9,188 observations (2,828 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 (asinh) 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 4 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 Table 4—women write 2–6 percent better than men.[55] 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 4 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. Thus, men's smaller effect potentially reflects their disproportionate tendency to co-author exclusively with other men—*i.e.*, precisely where the marginal impact of an additional woman is low.[56]

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

[55] 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.

[56] 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).

4.3 Comparing abstracts pre- and post-review.

In this section, I show that peer review causes (or at least exacerbates) the gender readability gap. To do so, I analyse papers before and after review by comparing published articles to their draft versions.

As discussed in Section 3, 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).

4.3.1 Summary statistics. Table 5 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).

Evidence from other fields indicates abstract text is altered during peer review (Hopewell et al., 2014).[57] Table 5 suggests this is similarly true in economics. The 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 and "hard" word to sentence count and (ii) hard word to word count. Between working paper and published versions, (i) decreases and (ii) increases (Table 5, second panel).[58] (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. Although hard *and* total word counts both decline, the latter falls 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 5, third panel), a finding consistent with similar investigations at medical journals (Biddle and Aker, 1996; Hayden, 2008; Roberts and Nolen-Hoeksema, 1994).[59] Thanks to fewer hard words per sentence, SMOG scores are higher in published articles regardless of gender. 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 is negative. In any case, women's papers endure comparatively

[57] Hopewell et al. (2014) compared 93 originally submitted manuscripts to their final versions published in *BMC*-series medical journals. Abstracts were altered in 16 percent of papers, generally because referees asked authors to tone down conclusions. I also collected 721 comments from ShitMyReviewer.com, a website dedicated to posting mean things written in referee reports. 180 (25 percent) deal with writing quality, document structure or word choice/tone. (Data and analysis available on request.)

[58] 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).

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

TABLE 5: 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); 272 published articles authored by at least 50 percent women (273 NBER working papers). Figures are means of textual characteristics by sex for NBER working papers and published articles. Last columns in each panel subtract working paper figures from published article figures for men (first panel) and women (second panel). Standard errors in parentheses. ***, ** and * difference statistically significant at 1%, 5% and 10%, respectively.

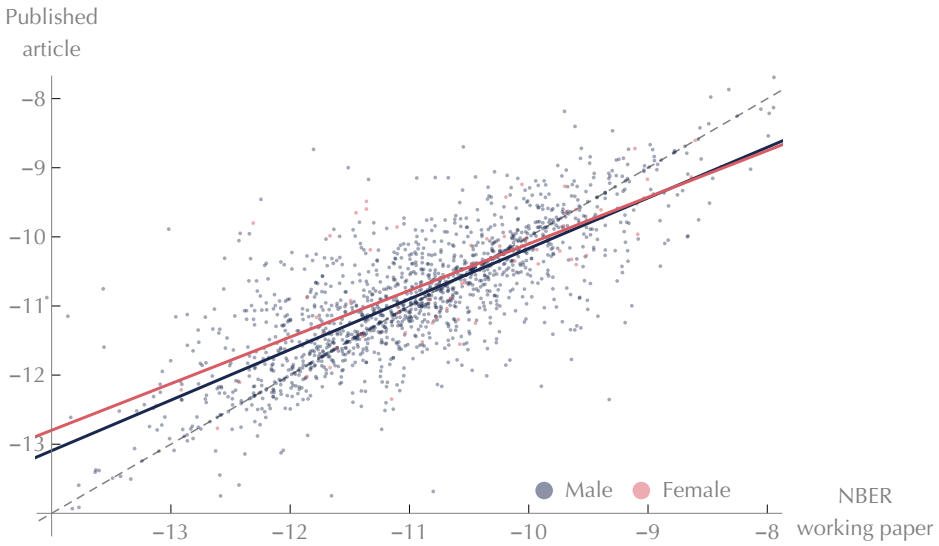

FIGURE 5: Published paper vs. draft readability

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.

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 5 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.

4.3.2 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 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.[60] ε_{jP} is P's error term.

[60] 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

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$_j$ 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):

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

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)).

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

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, *JEL* effects and max. T_j) and unobservable components, respectively.[61] 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.

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

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}:

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

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):

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

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.[62]

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}.[63] Roughly

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. Finally, Equation (3) implicitly controls for all factors that affect readability *before* peer review, *e.g.*, research field. Although certain fields may involve concepts that are easier to explain, their effect on readability is already present in the draft version of a manuscript; differencing scores eliminates it.

[61] 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.

[62] 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).

[63] 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.

restated, this implies that non-peer review factors are either independent of its timing (and therefore subsumed in version-invariant fixed effects) or unrelated to gender.[64] Section 4.3.4 evaluates this assumption.

4.3.3 Results. Table 6 presents results from OLS estimation of Equation (2) and Equation (3) and FGLS estimation of Equation (6) and Equation (7). Since gender bias is possible only when authors' identities are known or can be reasonably inferred, samples exclude 279 articles subjected to double-blind review before the internet; the following section considers their impact.[65] In an effort to maximise sample sizes, estimates in the first three columns omit field controls. Including them slightly increases standard errors; they otherwise make little difference (see Appendix L.2). Estimates in the final column implicitly account for field already (see Footnote 60).

The readability gap grew precisely while papers were being reviewed. Table 6's 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 L.1) are positive and significant—but only about 0.8. A less than unit value suggests μ_{jP} exerts downward pressure on R_{jW}'s coefficient, thereby artificially inflating first column figures.

Table 6's remaining columns present results from both strategies meant to deal with this bias. Columns two to four 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).[66] Column four displays their difference (β_{1P}); it is positive and significant for four out of five scores.

OLS estimates of β_{1P} from Equation (3) are shown in Table 6's final column. Their magnitudes and standard errors almost perfectly mirror FGLS estimates.

Consistent with evidence from Section 4.1, omitting field controls does not bias results. Table 6's final column implicitly accounts for any factor—including field—that affects readability before peer review (see Footnote 60). Although FGLS estimates could be biased by leaving *JEL* effects out, their resemblance to figures in the final column suggests not. Nevertheless, Appendix L.2 repeats the analysis with *JEL* effects. Results are similar to those presented here.

Double-blind review. Two journals—*QJE* and *AER*—employed double-blind review at some point during the time period covered by the data. *AER*'s spell began 1 July, 1989 and ended 1 July, 2011.[67] *QJE* used double-blind procedures until 1 June, 2005. *Econometrica* and *JPE* have never blinded referees to authors' identities.

[64] 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$_j$.

[65] Excluding these observations does not noticeably impact results or conclusions (for estimates based on the full sample, see Hengel, 2016, p. 18).

[66] FGLS difference (β_{1P}, column four) divided by the effect in published articles ($\widetilde{\beta}_{1W} + \beta_{1P}$, column three).

[67] From 1 May 1987 to 31 May 1989, half of the papers submitted to *AER* were evaluated by single-blind review; the remaining half were subjected to double-blind review (for details on the trial, see Blank, 1991). Referees correctly identified at least one author in 45.6 percent of double-blind reviewed papers—indicating that only about a quarter of the manuscripts were truly double-blind reviewed. I therefore classify every paper published during the trial as having undergone single-bind review. Excluding these observations from the analysis, however, has very little impact; estimated coefficients and standard errors are similar to those presented in Table 7 (results available on request).

TABLE 6: The impact of peer review on the gender readability gap

	OLS	FGLS			OLS
	Published article	Working paper	Published article	Difference	Change in score
Flesch Reading Ease	1.35**	2.27**	3.24***	0.97*	0.94
	(0.57)	(1.00)	(1.21)	(0.58)	(0.59)
Flesch-Kincaid	0.52***	0.32	0.77***	0.44**	0.44**
	(0.17)	(0.22)	(0.28)	(0.18)	(0.18)
Gunning Fog	0.52***	0.44*	0.86***	0.42**	0.41**
	(0.18)	(0.24)	(0.29)	(0.19)	(0.19)
SMOG	0.31**	0.32**	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**	0.12**
	(0.05)	(0.10)	(0.11)	(0.05)	(0.05)
Editor effects	✓	✓	✓		✓
Journal effects	✓	✓	✓		✓
Year effects	✓	✓	✓		
Journal×Year effects	✓	✓	✓		✓
Quality controls	✓[2]	✓[2]	✓[2]		✓[3]
Native speaker	✓	✓	✓		✓

Notes. Sample 1,709 NBER working papers; 1,707 published articles. Estimates exclude 279 pre-internet double-blind reviewed articles (see Footnote 65). Column one displays coefficients on female ratio (β_{1P}) from estimating Equation (2) directly via OLS (see Appendix L.1 for coefficients on R_{jW}); standard errors clustered by editor in parentheses. Columns two and three display $\widetilde{\beta}_{1W}$ and $\widetilde{\beta}_{1W} + \beta_{1P}$ from FGLS estimation of Equation (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 ✓[2] include citation count (asinh), max. T_j and max. t_j; ✓[3] includes max. t_j, only (see Footnote 60). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

To analyse the impact of double-blind review pre-internet, I re-estimate Equation (3) including Blind_j —a dummy variable equal to 1 if an article was subjected to double-blind review before Google incorporated in 1998—and its interaction with female ratio_j (Equation (8)).[68]

$$\begin{aligned} R_{jP} - R_{jW} = {} & \beta_{0P} + \beta_{1P}\,\text{female ratio}_j + \beta_{2P}\,\text{Blind}_j + \beta_{3P}\,\text{female ratio}_j \times \text{Blind}_j \\ & + \boldsymbol{\theta}_P \mathbf{X}_{jP} + \mu_{jP} + \varepsilon_{jP}. \end{aligned} \tag{8}$$

Table 7's first two rows display marginal effects of female ratio under non-blind (β_{1P}) and blind ($\beta_{1P}+\beta_{3P}$) review from OLS estimation of Equation (8). They suggest a smaller—possibly negative—gap under blinded peer review.[69] Marginal effects in single-blind reviewed papers (or those subjected to double-blind review post-internet) are identical to figures in Table 6.

Table 7's final row reports differences between effects (β_{3P}). Their consistent positive direction provides some (weak) evidence that masking authors' identities reduces peer review's impact on the gender readability gap. Nevertheless, (i) standard errors are large; and (ii) samples are

[68] Blind_j is equal to 1 for articles published during an official policy of double-blind review. A final publication date, however, may substantially lag the actual review date (for an illustration and discussion, see Blank, 1991). Because results are unchanged when including only *AER* articles published post May 1989 (see Footnote 67) and all *QJE* articles published before June 2005 were evaluated under double-blind review, misclassification errors are unlikely to substantially bias estimates presented in Table 7.

[69] In a preliminary version of this paper (Hengel, 2015), I estimated the impact of double-blind review on the gender readability gap in the sample of published papers, only. (It did not compare readability between versions as is done here.) It found that double-blind review corresponded to a higher readability gap. As shown in Appendix F, however, this conclusion is not robust to including the full set of publication year fixed effects.

TABLE 7: The impact of blinded peer review on the gender readability gap

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Non-blind	0.94	0.43**	0.41**	0.23**	0.13**
	(0.59)	(0.18)	(0.19)	(0.12)	(0.05)
Blind	−1.52	−0.56	−0.54	−0.36	−0.13
	(2.96)	(0.68)	(0.80)	(0.57)	(0.17)
Difference	2.45	1.00	0.95	0.60	0.26
	(3.05)	(0.73)	(0.84)	(0.59)	(0.17)
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Quality controls	✓[3]	✓[3]	✓[3]	✓[3]	✓[3]
Native speaker	✓	✓	✓	✓	✓

Notes. Sample 1,988 NBER working papers; 1,986 published articles. Columns displays the marginal effect on female ratio for papers undergoing non-blind (β_{1P}) and blind ($\beta_{1P} + \beta_{3P}$) review from OLS estimation of Equation (8). Standard errors clustered by year in parentheses. Quality controls denoted by ✓[3] include max. t_j, only (see Footnote 60). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

small—only 43 manuscripts in the blinded sample are authored by at least one woman; just 10 are exclusively female-authored.

In any case, whatever anti-bias effect double-blind review may have had *before* the internet disappeared *after* it. In Appendix F, I analyse the policy's post-internet impact. Gender differences are positive regardless of a journal's official review policy, suggesting that double-blind review is effective only as long as authors are not identifiable by other means.[70]

4.3.4 Robustness. Timing independence is the principle assumption required to causally link the readability gap to the peer review process. It is arguably only violated during the narrow timeframe after a manuscript is released as an NBER Working Paper but before it is submitted to a top-four journal.[71]

Just a small proportion of papers are exposed to this window. Figure 6 displays a histogram of the length of time between a working paper's release and submission to *Econometrica*. Most manuscripts are submitted to peer review at the same time or *before* they are released as NBER Working Papers. This is especially true of female-authored manuscripts.[72] Assuming similar submission-release patterns at *AER*, *JPE* and *QJE*, timing independence appears to be violated in only a small number of predominately male-authored papers.[73]

[70] In particular, double-blind review may be difficult to implement in academic fields with a culture of widely presenting, disseminating and publicising working paper results (*e.g.*, economics). Success is probably more likely in fields where pre-prints are private and review times are fast (*e.g.*, the physical sciences).

[71] That is, post-submission manuscript changes are probably only made inside peer review—either because referees actually request them or authors believe (possibly mistakenly—see Section 4.4) that they will be requested in a future revision.

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

[73] Additionally, most drafts have *already* been widely circulated prior to NBER Working Paper release. Average acknowledgment length in NBER Working Papers is 133 words. Most authors thank at least one person for comments—indeed, the vast majority thank several—and mention having previously presented the research in conferences and seminars. Combined with evidence from Figure 6, this suggests that gender differences in one's propensity to receive non-peer-review feedback only affects working paper readability and thus should not bias the results presented in Table 6.

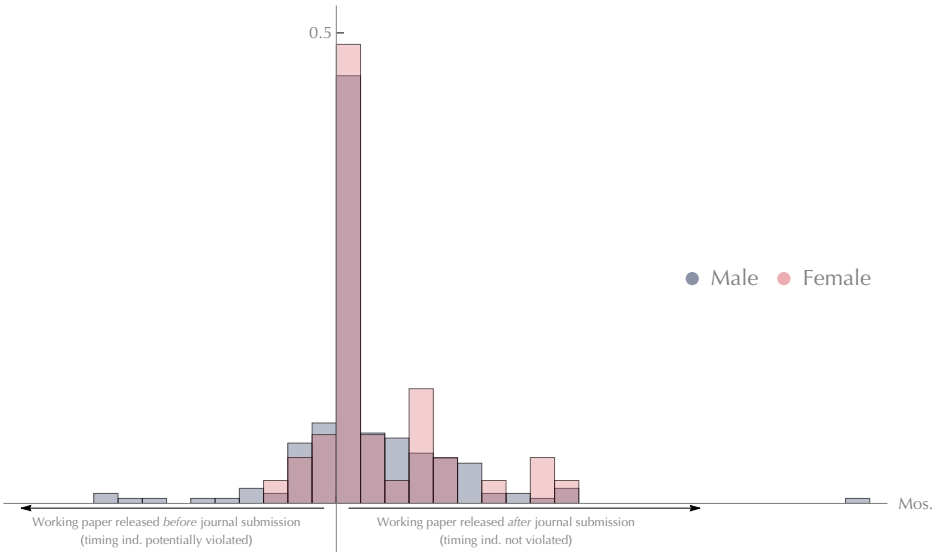

FIGURE 6: Distribution of months between NBER release and journal submission

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

Finally, NBER Working Papers have no abstract word limit; two journals in my sample do: *Econometrica* and *AER* restrict abstracts to 150 and 100 words, respectively. Gender differences in how authors conform to these limits may bias results in Table 6. To consider the possibility, I exclude the 642 observations—about 40 percent of the sample—with NBER abstracts longer than the official word limit of the journals in which they were eventually published. Results are presented in Appendix G. Coefficient magnitudes are similar to those in Table 6; standard errors are somewhat larger.

4.4 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).

4.4.1 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,

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

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

To aid the revision process, s writes a referee report from which i forms expectations about $\widetilde{R}^s_i$ by assigning subjective probabilities $\pi^s_{1it}(R)$ to all R. Unfortunately, the concept of readability is complex, some referees write insufficiently detailed reports and inattentive or hypersensitive authors misconstrue even perfectly clear advice. This renders i's interpretation of the report imprecise and his subsequent expectations about $\widetilde{R}^s_i$ 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.[74] Consequently, author i's stage 1 choice of R_{it} maximises his (immediate) subjective expected utility given s,

$$\Pi^s_{1it}(R_{it})u_i + \phi_{i|r_{0it}}(r_{1it}) - c_{i|r_{0it}}(r_{1it}). \tag{9}$$

$\Pi^s_{1it}(R_{it})$ is the cumulative sum of $\pi^s_{1it}(R)$ for all $R \leq R_{it}$; u_i is the utility of having a paper accepted in a prestigious journal;[75] $\phi_{i|r_{0it}}(r_{1it}) = \phi_i(R_{it}) - \phi_i(r_{0it})$ 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,

$$\int_\Sigma \Pi^s_{0it}(r_{0it}) v^s_{1it} \, \mathrm{d}\mu_i + \phi_i(r_{0it}) - c_i(r_{0it}), \tag{10}$$

where $\Pi^s_{0it}(r_{0it})$ is the cumulative sum for all $r \leq r_{0it}$ of author i's subjective probabilities $\pi^s_{0it}(r)$ about $\widetilde{r}^s_{0i}$; v^s_{1it} is Equation (9) 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

[74] 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.

[75] 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.

imply, at a minimum, that i updates Π^s_{0it} and Π^s_{1it} based on conclusive evidence derived from the choices and outcomes of equivalent peers (Definition 1)[76] and knowledge acquired during his own prior experience in peer review.[77]

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

Equation (9) and Equation (10) incorporate a variety of factors that potentially affect authors' readability choices—editorial standards ($\widetilde{r}^s_{0i}$ and $\widetilde{R}^s_i$); 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 (9) and Equation (10)—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 ($\widetilde{r}^s_{0i}$, $\widetilde{R}^s_i$, 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).[78] 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 D.

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,

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

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

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

[78] The analysis in Section 4.3 similarly establishes that (b) and/or (c) are significant factors driving the choice of R_{it}. It cannot, however, distinguish *between* (b) and (c).

or referee scrutiny is biased against i,

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

or both.

Theorem 1's three conditions are sufficient to verify discrimination 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 (9). As shown in Appendix D, $R^\star_i \leq r_{0it}$ 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 (10). 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

$$r_{0it} = \max\left\{R^\star_i, \widetilde{r}^{\overline{s}}_{0i} + e^{\overline{s}}_{0it}\right\} \quad \text{and} \quad R_{it} = \max\left\{r_{0it}, \widetilde{R}^s_i + e^s_{1it}\right\}, \tag{11}$$

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 $s \in \Sigma_{A_{it}}$.[79]

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.[80] 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 (11)). 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*

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

[79] As shown in Theorem 1's proof (Appendix D), 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.

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

where

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

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.[81]

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 D), 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.[82]

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 (13) does—providing a conservative estimate of D_{ik} under Theorem 1's weaker Condition 3.[83]

$$R_{it} - \max\{R_{it''}, R_{kt}\} \leq D_{ik}. \tag{13}$$

4.4.2 Empirical evidence. If topic, novelty and quality are appropriately controlled for, then discrimination is present when Theorem 1's three conditions hold at large enough t.

Consider first Condition 3—female-authored papers are accepted no more often than male-authored papers. The articles I evaluate have already been published, precluding gender analysis of acceptance rates. Nevertheless, the topic has been extensively studied elsewhere. Blank (1991) found that 12.7 and 10.6 percent of male- and female-authored papers were accepted at the *American Economic Review*, respectively.[84] 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

[81] 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.

[82] See also the discussion in Footnote 79 and Section 4.4.2.

[83] Although Equation (13) 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 (13)'s attenuation bias is much larger than the one generated by Equation (12).

[84] 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).

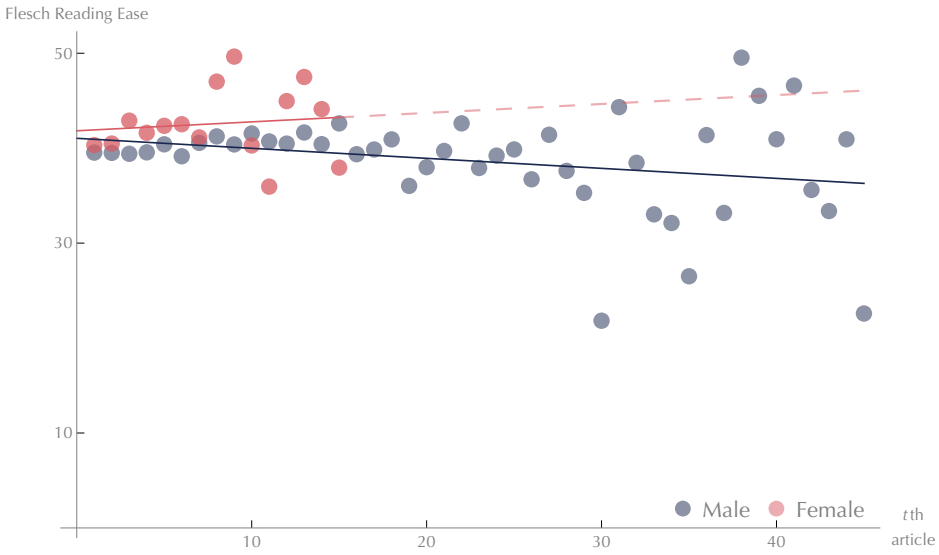

FIGURE 7: Readability of authors' tth publication

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).

33.3 percent, respectively (Gilbert et al., 1994).[85] There appear to be no gender differences in acceptance rates to NBER's Summer Institute (Chari and Goldsmith-Pinkham, 2017).[86] Desk rejection rates may actually be higher for female-authored papers submitted to the field journal *Energy Economics* (Gorelkina and Hengel, 2018; Tol, 2018). Ceci et al. (2014) provide a more comprehensive research review on the subject. Their conclusion: "When it comes to actual manuscripts submitted to actual journals, the evidence for gender fairness is unequivocal: there are no sex differences in acceptance rates." (Ceci et al., 2014, p. 111).

My data more cleanly identify Conditions 1 and 2. As their careers advance, women write more clearly: their average readability scores are 1–5 percent higher than the readability of their first papers, their latest papers 1–7 percent higher; for a man, however, his average and last paper are more poorly written than his first (Appendix H, Table H.1). Figure 7 suggests a similar story. It plots mean Flesch Reading Ease scores grouped by authors' tth article; as t increases, men's and women's readability diverges. Table H.2 and Table H.3 test the significance of that divergence, conditioning on numerous confounders. Once again, women write better; the magnitude of the difference increases as t increases.

Estimation strategy. Yet the set of women to satisfy one condition is conceivably orthogonal to sets that satisfy others; for Theorem 1 to apply, they must overlap. To address this concern, I 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

[85] These figures 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).

[86] 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.

to satisfy all conditions simultaneously; and (ii) generate (conservative) estimates of the effect of higher standards on authors' readability (Corollary 1).

Holding acceptance rates constant, Theorem 1 rules out confounding factors—*e.g.*, sensitivity to criticism and individual preferences—by comparing readability between equivalent authors experienced in peer review (Condition 1) and within authors before and after gaining that experience (Condition 2). I consider authors "experienced" by $t = 3$. 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 $\widetilde{R}_{i3}$; remaining errors are no longer gender specific: $e^s_{ni3} = e^s_{nk3}$, $n = 0, 1$ (Corollary 1).[87]

To account for equivalence, I match every female author with three or more publications (121) to her closest male counterpart (1,554). Matches were made using a Mahalanobis procedure with the following co-variates:[88] (1) T; (2) maximum citation count over t;[89] (3) institutional rank at $t = 1$;[90] (4) fraction of papers published per decade; (5) fraction of papers published by each journal; and (6) number of articles per primary *JEL* category. Co-variate balance pre- and post-match are shown in Appendix M.1. Appendix M.2 lists each matched pair. See Hengel (2017, pp. 30–33) for matches from a probit model performed with replacement and using a wider array of co-variates; results from alternative matching algorithms are available on request.

Under ideal circumstances, $R_{i3} - R_{i1}$ is the impact experience has on readability conditional on gender; $R_{i3} - R_{k3}$ measures gender's impact conditional on experience. Because of co-authoring, however, article gender is neither fixed over t conditional on i, nor is its difference constant between i and k, conditional on t. To account for this, I create a counterfactual $\widehat{R}_{it}$ that captures i's tth paper readability had it only been co-authored with members of i's same sex. It is reconstructed at female ratio equal to 1 for women and 0 for men using errors and coefficients from OLS estimation of Equation (14) in the gender and time appropriate subsample of authors:[91]

$$R_{it} = \alpha_{tg_i} + \beta_{tg_i} \text{female ratio}_{it} + \varepsilon_{it}, \tag{14}$$

where $g_i = m$ or $g_i = f$ if i is male or female, respectively. Regression output is shown in Appendix M.3. To adjust for the degrees of freedom lost when generating $\widehat{R}_{it}$, I inflate relevant standard errors by 1.05.

As long as individual authors' papers are roughly equivalent for all t, well-matched pairs account for remaining differences between i and k. They *don't* account for subtle variations over t conditional on i. Omitting factors exogenous to an author's long-term decision-making process—*e.g.*, year of publication or referee stereotypes about authors' institutions—potentially biases any estimate of D_{ik}.[92] But *including* t-varying factors under i's control—*e.g.*, journal, field and co-

[87] Recall that $e^s_{nit} - e^s_{nkt}$ converges to 0, so for large enough t Equation (12) and/or Equation (13) predict the direction of D_{ik} even when errors remain gender-specific.

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

[89] I eschew mean, median or minimum citation count in favour of the maximum on the assumption that an author's "quality" is principally a function of his best paper.

[90] Most people are at top ranked institutions by $t = 3$. By matching on $t = 1$ institution, I try to pair authors with similar career paths. In an earlier version of this paper, matches were based on highest institutional rank across all t; results are similar to those presented here (see Hengel, 2017, pp. 30–33).

[91] More specifically, I separately estimate Equation (14) in the following four subsamples: (i) female authors at $t = 1$; (ii) male authors at $t = 1$; (iii) female authors at $t = 3$; (iv) male authors at $t = 3$. I then generate $\widehat{R}_{it}$ using the appropriate coefficients and errors for each author: (i) $\widehat{R}_{i1} = \alpha_{1f} + \beta_{1f} + \varepsilon_{i1}$ for a female i at $t = 1$; (ii) $\widehat{R}_{i1} = \alpha_{1m} + \varepsilon_{i1}$ for a male i at $t = 1$; *etc.*

[92] Although, if referees harbour stereotypes about author's institutions in a way that is partially correlated with gender—*e.g.*, by holding institutions with more women to higher standards—then *unconstrained* women should adapt by moving to non-stereotyped institutions.

TABLE 8: $\underline{D}_{ik}$, 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	13.19	10.96	59	−7.75	7.85	23	4.85***	3.67***
							(1.12)	(1.21)
Flesch Kincaid	2.85	2.25	60	−2.50	2.41	23	0.94***	0.77***
							(0.26)	(0.27)
Gunning Fog	3.42	2.83	58	−2.58	2.73	23	1.20***	0.95***
							(0.30)	(0.33)
SMOG	2.74	1.98	51	−1.56	1.82	26	0.81***	0.63***
							(0.22)	(0.23)
Dale-Chall	1.38	0.92	63	−1.03	0.68	21	0.57***	0.48***
							(0.11)	(0.12)

Notes. Sample 121 matched pairs (110 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.

author characteristics—could too: the journals one submits to, the fields one gravitates toward and the people with whom one chooses to co-author are all endogenously determined by an author's experience. Higher standards may be met by manipulating any of these variables.

In spite of this, I don't actually find any evidence that observable t-varying factors—exogenous or not—drive women's increasing readability. For that reason, I opt for a parsimonious Equation (14). The robustness and validity of this approach are addressed in a following section.

Results. Table 8 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 (12)'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.[93] Male scores 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. In almost three-quarters of those, the member discriminated against was female.[94] Moreover, $\underline{D}_{ik}$ is (on average) 1.5 times as large (in absolute value) when discrimination is against women.

Figure 8 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 closer to zero. When women are discriminated against, $\underline{D}_{ik}$ is more spread out. Furthermore, instances of obvious discrimination are predominately against women: $\underline{D}_{ik}$ is five times more likely to be one standard deviation above zero than below it.

[93] 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 M.1 and the next section for a discussion).

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

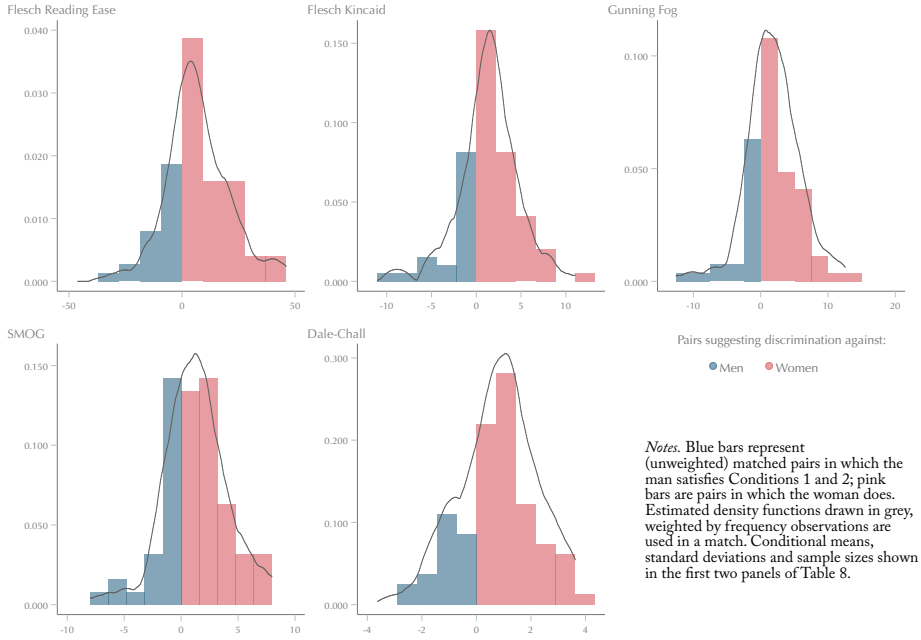

FIGURE 8: Distributions of $\underline{D}_{ik}$, Equation (12)

Table 8'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.[95] Mean $\underline{D}_{ik}$ is positive and significant in both columns for all five scores. First column figures suggest that higher standards cause senior female economists to write (at least) seven percent more clearly than they otherwise would.[96]

Appendix M.4 replicates Table 8 using Equation (13) to estimate D_{ik}. By definition, figures are smaller. Conclusions, however, are identical.

Robustness. Conclusions drawn from Table 8 are principally predicated on three assumptions: (1) i and k are equivalent; (2) 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;[97] (3) experience is the only exogenous t-varying factor driving within i changes in readability, conditional on a paper's share of female authors. If any is violated, discrimination against women cannot be inferred from an overrepresentation of matched pairs with $\underline{D}_{ik} > 0$.

Assumption (1) depends on match accuracy. Post-match co-variates are well balanced (Appendix M.1). They remain well balanced—and similar to the matched population—when restricted to pairs satisfying $\underline{D}_{ik} \neq 0$. To facilitate further scrutiny, Appendix M.2 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 generate them. Outcomes, however, are not. After controlling for T, decade, journal

[95] 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.

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

[97] 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.

and *JEL* code, matches using alternative variables (*e.g.*, minimum citation counts and mean institutional rank) and specifications (*e.g.*, propensity score matching) generate similar figures and conclusions (available on request).[98]

Assumption (2) 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}. Fifty percent of women with three or more top publications satisfy Conditions 1 *and* 2 when compared to equivalent men.[99] Among them, $\underline{D}_{ik}$ is far from zero: these women write, on average, 21 percent more clearly than equivalent men with identical experience. 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 $e^s_{ni3} = e^s_{nk3}$.[100] When violated, I can no longer conclude that $\underline{D}_{ik}$ conservatively estimates D_{ik}.[101] 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 $\underline{D}_{ik}$.[102]

As for assumption (3), I find no evidence that exogenous *or* endogenous observable t-varying factors drive women's increasing readability. Appendix H and additional analyses in a 2016 version of this paper (Hengel, 2016, pp. 23–24) show an identical pattern despite controlling for a large array of potential confounders. In a 2017 version, I reconstructed $\widehat{R}_{it}$ using several t-varying factors (number of co-authors, institutional rank, institutional rank of the highest ranked co-author, t for the most experienced co-author, publication year and dummies for each journal) (Hengel, 2017, pp. 30, 61); Appendix M.5 adds *JEL* classification codes to Equation (14). In Table J.15 (Appendix J.3) I restrict Table H.2's analysis to solo-authored papers or those co-authored by members of the same sex.[103] In all instances, women's readability is consistently shown to increase with t; when comparable results are estimated, they are similar to those presented in Table 8 and Figure 8.[104]

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) $r_{0k3} \leq r_{0i3}$—

[98] In an earlier version of this paper, I matched authors using propensity scores from a probit model with replacement. See Hengel (2017, Section 3.4.3, pp. 30–34) for results.

[99] Women are the better writers in 69 percent of matched pairs. In 30 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.

[100] $\underline{D}_{ik}$ actually remains a causal, conservative estimate of the impact of discrimination 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 D.

[101] Specifically, this assumption is violated if at $t = 3$ the women listed in Appendix M.2 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 $t = 3$.

[102] For a description of this downward bias, see the discussion on Corollary 1 in Section 4.4.1 and its proof in Appendix D.

[103] To test whether co-author characteristics—regardless of exogeneity—are behind women's increasing readability, I would ideally repeat the analysis on the subset of matched pairs in which the woman's $t = 1$ *and* $t = 3$ papers are solo- or exclusively female-authored. This criteria is satisfied in only 16 matched pairs. Nevertheless, a roughly similar proportion satisfy $\underline{D}_{ik} \neq 0$; in most of those, $\underline{D}_{ik} > 0$, suggesting discrimination against the female member. In 20 matched pairs, the woman's $t = 1$ and $t = 3$ papers are majority female-authored; in 76, they're at least 50 percent female-authored. In both restricted samples, $\underline{D}_{ik} \neq 0$ in about 60–70 percent of matched pairs; most of those (70–80 percent) indicate discrimination against the female member. Discrimination's average impact across all five scores is also about 7 percent.

[104] See also Footnote 141 for evidence suggesting women are *not* more likely to co-author with other women as t increases.

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.[105]

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, I reviewed the literature on gender neutrality in journals' acceptance rates earlier in this section. To recap, women are not accepted more often than men. In Appendix M.4, I attempt to control for acceptance rates 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 4.3, women's draft papers are indeed more readable than men's. Section 4.5 provides further confirmation. Figure 9 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 $t = 1$.

4.5 Direct vs. indirect effect of higher standards

Women can respond to higher standards in peer review immediately (direct effect) or pre-emptively (indirect effect). A bank-of-the-envelope calculation based on analyses in Section 4.3 and Section 4.4.2 suggests the direct effect is roughly responsible for a fifth to a fourth of the gender readability gap caused by discrimination.[106]

Yet the weight of each effect likely depends on authors' information about—hence experience with—the peer review process. To illustrate the evolution of their relative importance, Figure 9 compares papers pre- and post-review as authors' publication counts rise. Hollow circles denote NBER draft readability; solid diamonds reflect readability in the final, published versions of those same papers; dashed lines trace readability as papers undergo peer review. Numbers derived from FGLS estimation of Equation (15):

$$R_{j_{it}m} = \beta_0 + \beta_1 \text{ female ratio}_j + \beta_2 \text{ female ratio}_j \times t_i + \beta_3 t_i + \boldsymbol{\theta} \mathbf{X}_j + \varepsilon_j, \qquad (15)$$

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 (asinh) 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 4.2).[107] The first panel of Table 9 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; estimates correspond to the lengths of the dotted lines in Figure 9. Panel two shows the marginal effect of female ratio (β_2) for each version of a manuscript. Differences are shown in the final row.

[105] That is, women's papers are desk rejected at higher rates *because* their drafts are more poorly written. This does not preclude situations in which women's drafts are more likely desk rejected *despite* being better (or equally well) written.

[106] Results in Table 6, columns four and five (direct effect) divided by those in Table 8, columns (1) and (2) (sum total of direct and indirect effects), averaged over scores.

[107] 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.

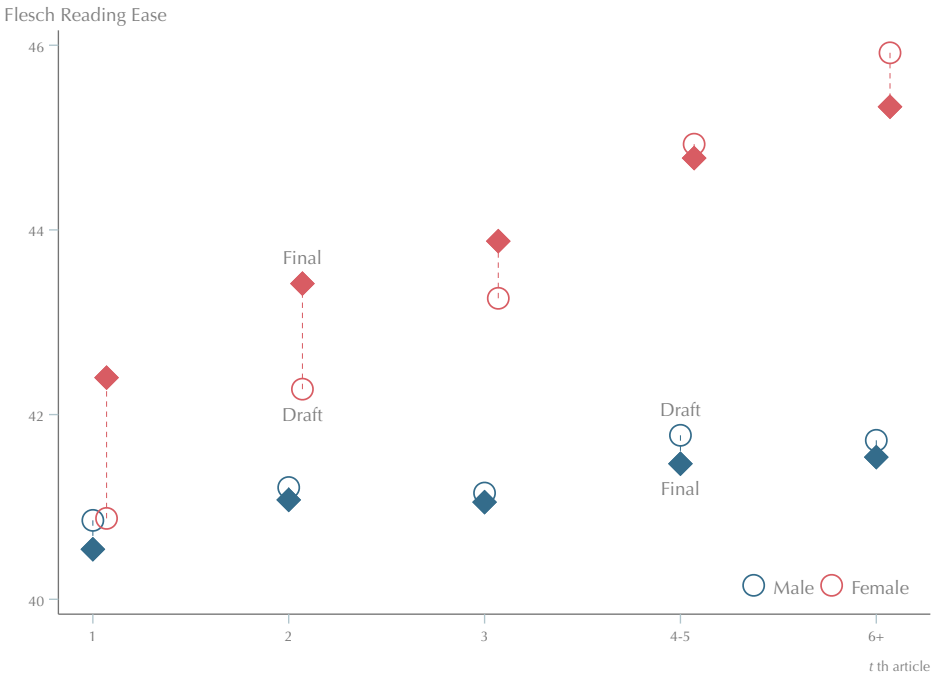

FIGURE 9: Readability of authors' tth publication (draft and final)

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. Hollow circles denote estimated readability of NBER working papers from FGLS estimation of Equation (15); solid diamonds show the estimated readability in published versions of the same papers. Controls are: editor, journal, year, journal and year interactions, English fluency dummies and quality controls (citation count (asinh) 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.

For $t = 1$ and $t = 2$, differences are large, positive and significant; three and up, they're fairly small. The readability gap in the published article is statistically significant and relatively stable at every t (Table 9, second panel). Increasingly, however, it forms before submission. Draft readability contributes nothing to the gap at $t = 1$. That rises to 40 percent at $t = 2$ and 70 percent at $t = 3$. By $t = 4$–5 and $t = 6+$, men and women mostly address referee concerns before peer review.

A number of tentative conjectures relevant to the gender readability gap are apparent from Figure 9 and Table 9. First, the direct effect dominates when women have less experience; the indirect effect dominates when women have more experience. This pattern of behaviour implies that women initially underestimate referees' thresholds but learn about them over time and adapt their *ex ante* writing style accordingly.

Second, inexperienced men and women seem to make similar choices in draft readability.[108] This suggests identical initial preferences for and beliefs about the importance of writing well. In one important sense, however, men are still better informed—the standards they believe apply actually do; junior women appear to mistakenly assume similar standards apply to them, too.

Third, experienced men *and* women sacrifice time upfront to increase acceptance rates and/or speed up review. All things equal, economists who anticipate referees' demands are rejected less often and/or peer reviewed faster. The cost is spending more time revising a manuscript before submitting it. Assuming choices by senior economists express optimal tradeoffs with full information, Figure 9 and Table 9 imply little—if any—gender difference in these preferences.

[108] Point estimates are almost indistinguishable, but the standard error of their difference is relatively large (see Table 9); please interpret this result with caution.

TABLE 9: Readability of authors' tth publication (draft and final)

	$t=1$	$t=2$	$t=3$	$t=4$–5	$t\geq 6$
Predicted $R_{jP}-R_{jW}$					
Women	1.52**	1.14*	0.62	−0.15	−0.58
	(0.64)	(0.60)	(0.70)	(0.90)	(1.17)
Men	−0.31*	−0.13	−0.10	−0.31**	−0.18
	(0.18)	(0.11)	(0.09)	(0.15)	(0.20)
Marginal effect of female ratio					
Published article	1.86*	2.34***	2.83***	3.31***	3.79**
	(1.03)	(0.74)	(0.75)	(1.06)	(1.48)
Draft paper	0.02	1.07	2.11**	3.15***	4.20***
	(1.21)	(0.92)	(0.82)	(0.96)	(1.26)
Difference	1.84***	1.28*	0.72	0.16	−0.40
	(0.70)	(0.67)	(0.79)	(1.00)	(1.27)
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Quality controls	✓[5]	✓[5]	✓[5]	✓[5]	✓[5]
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 direct 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 ($\beta_1+\beta_2\times t$), separately for draft papers and published articles. Figures from FGLS estimation of Equation (15). Quality controls denoted by ✓[5] include citation count (asinh) 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.

Finally, accepted papers by inexperienced female economists appear to sustain the toughest review. Assuming the same is true of rejected manuscripts, Figure 9 and Table 9 indicate higher desk rejection rates for junior women's work.[109] This seems to be true at the field journal *Energy Economics* (Gorelkina and Hengel, 2018; Tol, 2018).[110] Unfortunately, I cannot test this hypothesis with the data I have collected. In Section 4.6, however, I investigate the former claim, *i.e.*, that accepted manuscripts by junior women (or at least women) undergo the toughest—hence longest—review.

4.6 Duration of peer review

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

Higher standards for female authors probably means their papers spend longer under review, all else equal. 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). And an essay's rhetorical competency is highly correlated with the length of time one is given to compose it (Hartvigsen, 1981; Kroll, 1990).

To investigate, I turn to *Econometrica*, the only journal among the four to make disaggregated data on the revision process publicly available. Figure 10 is a histogram of time (in months)

[109] Higher desk rejection rates and tougher review are substitutes. Tougher review probably dominates when editors (or even referees) monitor and implement a policy of gender neutral acceptance rates. Otherwise, placing higher standards on female-authored work likely results in higher desk rejection rates (or some combination of both).

[110] Gorelkina and Hengel (2018) also finds that accepted papers with a female corresponding author spend 27–29 days longer in peer review, go through more revision rounds and receive more referee reports.

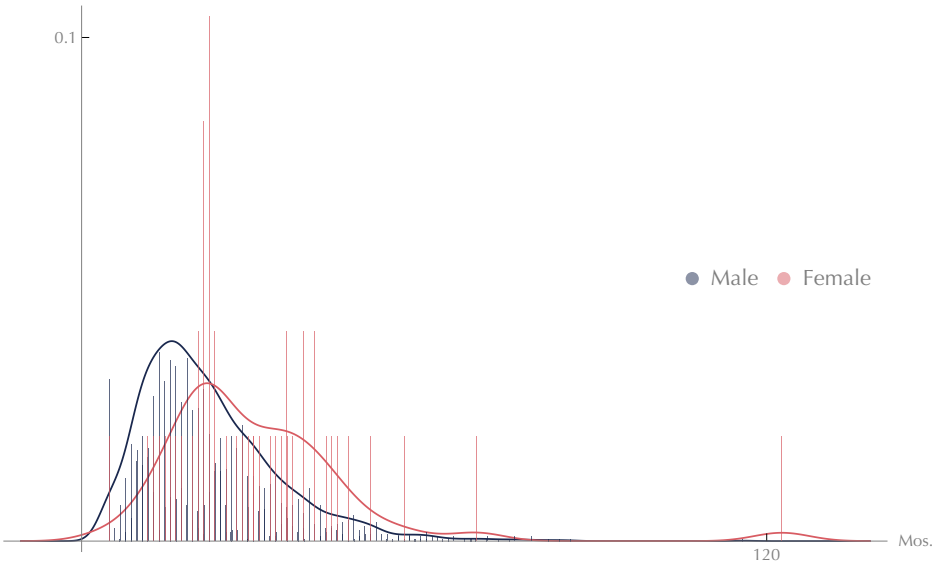

FIGURE 10: Distribution of review times at *Econometrica*

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 (2,398); pink bars are papers written only by women (48). Source: *Econometrica*.

between dates papers are first submitted to and their final revisions received by *Econometrica*'s editorial office.

Since 1950, *Econometrica* has published 52 papers authored entirely by women.[111] Their review times disproportionately cluster in Figure 10'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. Of the four articles in the entire sample with the longest review times, three are authored by at least one woman.[112] Among papers published post–2005, 27 percent in the top quartile have at least one female co-author.[113]

For more precision on gender differences in review times, I build on a model by Ellison (2002, Table 6, p. 963) and estimate Equation (16):

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

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,[114] $\max t_j$ 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 published 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.[115]

[111] Submit-accept times were not available for four of these articles. 180 mixed gender co-authored papers are not included in Figure 10; they *are* included in Table 10's estimation sample.

[112] The only one published before 2010 is "Dynamic Aspects of Earning Mobility" by Lee A. Lillard and Robert J. Willis (both men). It was submitted in October 1967, accepted in April 1977 and published in September 1978.

[113] For comparison, only 19 percent of all papers published post–2005 are co-authored by at least one woman.

[114] If one co-author goes on maternity leave or has young children, I assume another co-author manages the revision

TABLE 10: Revision duration at *Econometrica*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female ratio	5.411**	6.805***	6.802***	5.801***	6.822***	9.037***	9.079***
	(2.021)	(2.100)	(2.087)	(2.060)	(2.092)	(2.630)	(2.560)
Max. t_j	−0.175***	−0.180***	−0.177***	−0.176***	−0.175***	−0.178**	−0.184**
	(0.061)	(0.062)	(0.062)	(0.062)	(0.062)	(0.075)	(0.075)
No. pages	0.197***	0.196***	0.195***	0.196***	0.195***	0.233***	0.219***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.040)	(0.041)
N	1.114**	1.070**	1.060**	1.094**	1.067**	1.353**	1.243*
	(0.423)	(0.420)	(0.422)	(0.423)	(0.422)	(0.601)	(0.638)
Order	0.219**	0.216**	0.214**	0.217**	0.214**	0.477**	0.461**
	(0.086)	(0.086)	(0.086)	(0.086)	(0.087)	(0.190)	(0.198)
No. citations (asinh)	−0.335	−0.359*	−0.347*	−0.333	−0.358*	−0.555	−0.569
	(0.204)	(0.203)	(0.203)	(0.204)	(0.202)	(0.493)	(0.495)
Mother			−7.223***		−11.353***	−19.832***	−19.847***
			(2.597)		(3.083)	(3.947)	(3.678)
Birth				−3.468	6.895*	14.207**	13.946**
				(3.280)	(4.084)	(5.746)	(5.539)
Editor effects	✓	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓	✓	✓
JEL (primary) effects							✓
No. observations	2,625	2,610	2,625	2,625	2,625	1,281	1,281

Notes. Sample 2,626 articles. Coefficients from OLS estimation of Equation (16); (2) excludes papers authored only by women who gave birth (9 articles) and/or had a child younger than five (15 articles) at some point during peer review; (6) and (7) exclude papers published before 1990. Standard errors clustered by year in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Table 10 displays results across a range of specifications. All models include editor, year and institution fixed effects. Year effects refer to the year in which an article was published; see Appendix N.1 for results using years of submission and acceptance, instead. Column (1) does not control for motherhood or childbirth; (2) drops papers authored entirely by women who had children younger than five and/or gave birth during peer review; (3) controls for motherhood but not childbirth; (4) controls for childbirth but not motherhood; (5) controls for both childbirth and motherhood; (6) and (7) restrict the sample to papers published after 1990; (7) includes fixed effects for primary *JEL* categories.[116]

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.[117] The average male-authored paper takes 18.5 months to complete all revisions; papers by women need more than half a year longer.[118]

Why? Well, it's not motherhood. Yes, giving birth slows down review—responding to referees is apparently put on hold for the first 6–12 months of a newborn's life—but having a young child

process unless she, too, faces similar family commitments.

[115] 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*.

[116] *JEL* classifications are only available for papers published after 1990 (see Section 3); Table 10's column (7) estimates Equation (16) on about half of the data. Column (6) is likewise estimated on the truncated sample but excludes *JEL* effects.

[117] This conclusion is robust to altering the age-threshold on mother_j (see Appendix N.2).

[118] Based on results in (5). Male effect estimated with zero female co-authors. When publication year fixed effects are replaced with submission and acceptance year fixed effects, female-authored papers spend 3–4 months and 6–9 months longer in peer review, respectively (see Appendix N.1).

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?[119] 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.[120] 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 15 articles in the sample.[121]

As for Table 10'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.[122]

4.6.1 Junior women undergo the longest review. Figure 11 tests Section 4.5's finding that junior women undergo the toughest peer review. Blue dots are coefficients on female ratio from FGLS estimation of Equation (16) on sub-samples of junior ($t_i = 1$) and senior ($t_i > 1$) authors;[123] the yellow dot is their difference. To eliminate confounding by more senior co-authors, I restrict the sample to authors satisfying max. $t_j = t_i$.[124] 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 4.2 for details).[125]

Figure 11 suggests papers by junior women take longer in review. When combined with evidence from Section 4.5, it supports the hypothesis that senior women have an easier time *in* peer review only because they address higher standards *before* it. Nevertheless, few junior women—and even fewer senior women—publish in *Econometrica*; results are sensitive to specific observations.

In the present paper, I do not explore the particular relationship between review time and readability. Superficial correlations between the two are almost certainly confounded by unobserved heterogeneity in submission quality. For example, if "referees feel the need to demonstrate their intelligence or industriousness to editors by identifying problems in [otherwise perfectly fine] papers", they may "inflate minor blemishes"—*e.g.*, readability—"to the status of major flaws" (Berk et al., 2017, p. 231). It is undoubtedly easier to remedy most issues surrounding readability than it is to fix actual, major flaws. Thus, the unbiased relationship between readability and review time is impossible to obtain without controlling for submission quality or placing significant structure on the data generating process. This is beyond the scope of the present paper.[126]

[119] Ginther and Kahn (2004) find evidence of this hypothesis in a survey of doctoral recipients ten years after receiving their Ph.D. They find that women with children are more productive than all men and women without children (although the difference is not significant). They suggest that less productive women may decide not to have children.

[120] 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—I (perhaps unfairly) give this interpretation less weight.

[121] The count increases to 16 and 18 articles when mother_j's threshold is defined as children younger than ten and 18, respectively (see Appendix N.2).

[122] 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.

[123] In Section 4.4.2, I define "experienced" as $t = 3$. However, most female-authored papers published in *Econometrica* are by women with no (or only one) previous top publication; only 17 women in the relevant sample had two or more.

[124] When the sample includes all authors, standard errors are larger and effect sizes (and their differences) much smaller (regression output available on request).

[125] Results and conclusions based on unweighted regressions are very similar to those presented here (output available on request).

[126] Unobserved heterogeneity in submission quality should not bias conclusions elsewhere in the paper as long as it does

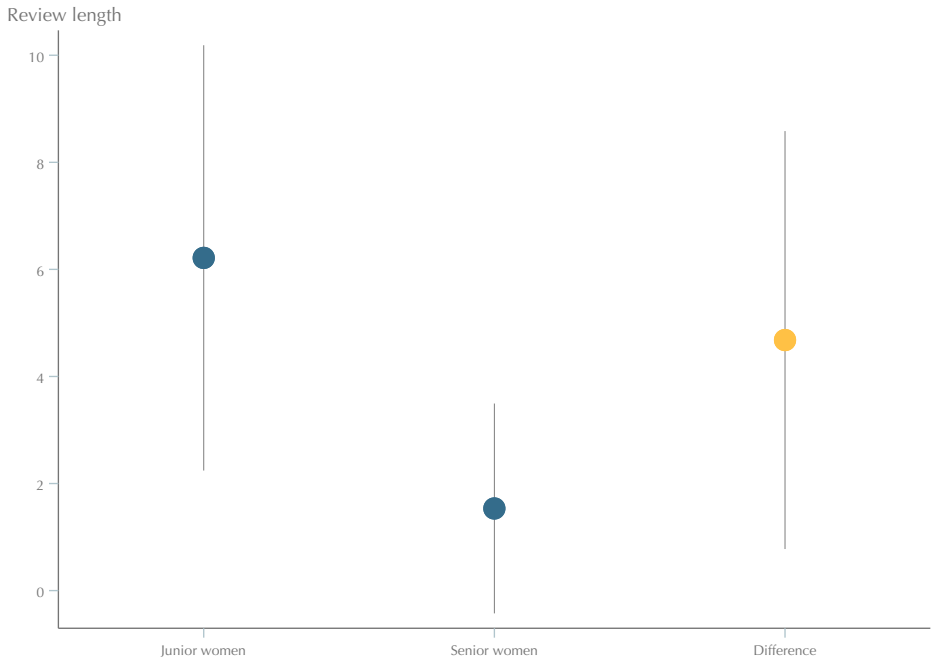

FIGURE 11: Coefficient on female ratio, junior vs. senior women

Notes. Sample 3,019 observations satisfying max. $t_j = t_i$; 1,925 distinct articles and 1,996 distinct authors (81 female). Blue dots are the coefficients on female ratio corresponding to separate FGLS estimations of Equation (13) on authors for whom $t_i = 1$ (junior) and $t_i > 1$ (senior), respectively. The yellow dot is their difference. Regression weighted by $1/N_j$.

5 Summary and discussion

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

Why? Either women voluntarily improve their writing during peer review—*e.g.*, because they're more sensitive to criticism—or better written papers are women's response to higher standards imposed by referees and/or editors.

To theoretically distinguish between hypotheses, I construct a simple dynamic model of an author's decision-making process. Its intuition is straightforward. Assuming preferences are fixed over time, authors improve readability today relative to yesterday only when they believe more readable papers are more often accepted. And while oversensitivity and/or poor information may distort their beliefs—and in turn affect readability—with experience, they correct those mistakes. Thus, when an experienced author writes more clearly than her inexperienced self, the change improves her acceptance rate. If that same experienced author writes more clearly than an equivalent, experienced author of the opposite gender, *yet* her papers are no more likely to be accepted, *then* asymmetric editorial standards and/or biased referee assignment—*i.e.*, discrimination—explains the difference.

not partially correlate with the share of female authors on a manuscript. If women are held to higher standards, however, this assumption is violated when controlling for review time and estimating the impact gender has on readability (or visa versa). Conditioning on one when estimating the other opens up a backdoor channel to the unobserved heterogeneity. Although controlling for citations introduces a similar concern, Figure 2 and Table 10 suggest correlations are weak (and hence bias low). Hengel (2016) presents most analyses without citation controls; results are very similar to those found here.

The model establishes three conditions sufficient to demonstrate higher standards are present in academic review: (i) experienced women write better than equivalent men; (ii) women improve their writing over time; and (iii) female-authored papers are accepted no more often than equivalent male-authored papers. I empirically test (i) and (ii) by matching experienced, well-published female economists to similarly productive men on observables determining the topic, novelty and quality of their research. The resulting estimate suggests discrimination causes experienced women to write at least seven percent more clearly than they otherwise would.[127]

5.1 Evaluating alternative explanations

A gender readability gap exists. It's still there after including editor, journal and year effects—meaning it's hard blame specific policies or attitudes in the fifties, long since overcome. The gap is unaffected by field controls—*i.e.*, it unlikely results from women researching topics that are easier to explain. Nor does it appear to be caused by factors correlated with gender but actually linked to authors' (or co-authors') competence as economists and fluency in English—if so, institution and native speaker dummies would reduce it. They do not.[128]

The gap grows between first draft and final publication and over the course of women's careers, precluding inborn advantage and one-off improvements in response to external circumstances unrelated to peer review. This likewise rules out gender differences in (i) biology/behaviour—*e.g.*, sensitivity to referee criticism[129]—or (ii) knowledge about referee expectations. If diligently addressing every referee concern has no apparent upside—acceptance rates are unaffected—and a very clear downside—constant redrafting takes time—even the most oversensitive, ill-informed woman would *eventually* re-examine initial beliefs and start acting like a man.[130] Yet this is not what we observe. The largest investments in writing well are made by female economists with greatest exposure to peer review—*i.e.*, those with the best opportunity to update their priors.

Women's papers are more likely assigned female referees (Abrevaya and Hamermesh, 2012; Gilbert et al., 1994).[131] If women are more demanding critics, clearer writing could reflect their tougher reviews.[132] Women concentrate in particular fields, so it's natural female referees more often review female-authored papers. Nevertheless, for the readability gap to exist only because of specialisation, controlling for *JEL* classification should explain it.[133] It does not: including 20 primary or 718 tertiary *JEL* category dummies has little effect. So if referee assignment is causing the gap, it's only because journals disproportionately refer female-authored papers to the toughest

[127] This conclusion assumes male-authored papers are accepted at least as often as female-authored papers, which cannot be tested using the data I have collected. See Section 4.4.2 for prior research justifying this claim.

[128] 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.

[129] 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).

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

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

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

[133] 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.

critics.[134] Meaning it isn't referees who are biased—it's editors.[135]

Section 4.3 directly links an increase in the gender readability gap to peer review; Section 4.4 establishes that factors outside women's control—assumed, at this point, entirely peer-review-related—drive it. Yet oversensitivity and/or poor information could create the former gap while *another* gender bias unconnected to peer review generates the latter. One in particular comes to mind: the feedback women receive in conferences and seminars. Perhaps experienced female economists tighten prose (before or after submission) in response to audience member remarks.

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

In the universe of straightforward alternatives, this leaves us with one: female economists are less capable researchers. As mentioned earlier, factors correlated with gender but actually related to competency should decline when appropriate proxies are included. The sample itself is one such proxy—these are, after all, only articles published in the top four economics journals. Adding other controls—author institution, total article count, citation counts and published order in an issue—has no effect.[136] 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 4.4.2 are not really equivalent.[137] (To decide for yourself, see Appendix M.2.) And if this is true, editors and referees *should* select and peruse our papers more carefully—a byproduct of which could be better written papers after-the-fact or more attractive prose compensating for structural weaknesses before it.[138]

"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

[134] Relatedly, perhaps female-authored research is more provocative and therefore warrants more scrutiny. Yet if this explained the gap, controlling for *JEL* classification should reduce (or eliminate) it—unless women's work is systematically more provocative even among researchers in very narrow fields. There is some evidence for this hypothesis—provocative work is (presumably) highly cited work and recent female-authored papers published in top economics journals are cited more (Hengel, 2018). Yet more provocative, cited research would probably be published at higher rates—and there is no evidence women's paper's are more frequently accepted (Ceci et al., 2014). In any case, women respond to incentives just like men; if we could get boring papers published, we'd write them.

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

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

[137] If readability is an unbiased proxy for "quality", then my results suggest women's papers are better. Otherwise, the gender gap only reflects marginal differences in quality on one narrow dimension. But as long as readability does not substitute for lower quality elsewhere conditional on observables (including subsequent citations), a positive gap suggests higher standards for women's writing.

[138] It does seem contradictory, however, that women would be capable of writing better than men—even before referee input (Table 6)—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. 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.

Section 4.4.1, gender has very little impact on the latter;[139] a review of past studies on male vs. female citations find four in which women's papers received fewer, six where they were cited more and eight with no significant difference (Ceci et al., 2014). Recent research specific to economics suggests contemporary female-authored papers get cited more (Grossbard et al., 2018; Hengel, 2018).

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?[140] Maybe women actually enjoy being poorly informed, overconfident and sensitive to criticism—or (more likely) I could have otherwise misspecified the author's objective function in Section 4.4.1. Meanwhile, a preference for writing well coupled with unaccounted for exogenously determined co-author characteristics could combine to cause women's more readable papers *and* their increasing readability[141]—although restricting the analysis to solo-authored papers or those co-authored by members of the same sex results in similar figures and identical conclusions (see Appendix J.3, Table J.15 and the robustness discussion in Section 4.4.2).[142] Alternatively, measurement error and/or co-variate controls could have interacted with gender in ways I did not anticipate.[143] And of course, the statistically significant relationships this paper documents may simply be unfortunate (particularly for me!) flukes.[144]

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

Moreover, several studies document a gender difference in critical feedback of similar form—employee performance reviews and student evaluations. Ongoing research suggests female workers are held to higher standards in job assessments. They are acknowledged less for creativity and technical expertise, their contributions are infrequently connected to business outcomes; guidance or praise supervisors do offer is vague (Correll and Simard, 2016).[145]

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 be-

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

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

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

[142] Relatedly, women may have preferred to have written their $t = 1$ publication more clearly, but senior male co-authors held them back; at $t = 3$, they enjoy more freedom to achieve their desired (higher) readability by writing on their own or with other women. This runs counter to the observation in Footnote 141, however. Moreover, women are more likely to co-author with more senior men at $t = 3$ than they were at $t = 1$.

[143] Section 2.1 outlines principle sources of measurement error as well as steps I have taken to minimise their impact. Meanwhile, coefficient magnitude and standard errors remain relatively stable when gradually introducing controls (Table 3), reducing the likelihood of "collider" bias (see Footnote 64).

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

[145] 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).

ing "smart", "humble" or "cool" (Schmidt, 2015).[146] 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).

5.2 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.[147]

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.[148]

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.[149] 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)[150]—but if more accountability promotes fairer outcomes, ethical arguments in its favour should outweigh minor practical concerns.[151]

6 Conclusion

Female-authored articles published in top economics journals are better written than similar papers by men. A large part of this gap originates in peer review. To identify the causal mechanism, I derive an empirical test from a subjective expected utility model of an author's decision-making

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

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

[148] I find weak evidence suggesting the policy may have been effective before the late 1990s (Section 4.3) but not afterwards (Appendix F).

[149] 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.

[150] 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).

[151] A decision like this should be carefully considered based on a body of evidence and ideally extensively tested before officially rolling out.

process. The resulting estimate suggests higher standards in peer review cause senior female economists to write at least seven percent more clearly than they otherwise would.

I also document evidence that sheds more light on the impact higher standards have on women's choices and their productivity measurement: (i) female-authored papers spend more time under review; (ii) women appear to quickly adapt to higher standards *in* peer review by adjusting their behaviour *before* it. The latter emphasises that observational data almost always capture discrimination in equilibrium; to identify it, one must determine where the equilibrium would have been had discrimination not occurred. The former suggests higher standards could play a role in academia's "Publishing Paradox".

If higher standards apply elsewhere in the economy they may also rationalise many instances of lower 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 measurement of female productivity—billable hours and client revenue decline; female lawyers appear less productive than they truly are.

Finally, this study exploits publicly available data to identify and evaluate discrimination's impact on those who are discriminated against. But it is narrowly focused and—because it doesn't study the bias's source—proposes few solutions. More research is needed to establish external validity, be confident in internal validity and better understand how and where higher standards for female authors emerge in peer review. Future studies could apply a similar theoretical framework and more holistic measure of research quality—*e.g.*, citations—to these and other economics journals. Readability scores analysed using a similar setup could also 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.* But to answer the deeper questions this paper raises requires confidential data on the editorial process—including access to referee reports. I hope journals are challenged to be candid about bias and willing to support the access and research needed to better understand it.

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.

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

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 (2018). "Equal but Inequitable: Who Benefits from Gender-Neutral Tenure Clock Stopping Policies?" *American Economic Review* (forthcoming).

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.

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

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.

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

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 (2018). "Do Women Ask?" *Industrial Relations: A Journal of Economy and Society* (forthcoming).

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.

Ayres, I. (2001). *Pervasive Prejudice? Unconventional Evidence of Race and Gender Discrimination*. Chicago, Illinois: University of Chicago Press.

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.

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

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

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.

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.

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

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.

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

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 (2017). "The Gender Wage Gap: Extent, Trends, and Explanations". *Journal of Economic Literature* 55 (3), pp. 789–865.

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.
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., A. Freeman, and B. Levy (1983). "Minimum Competency Testing of Reading: An Analysis of Eight Tests Designed for Grade 11". In: *The Courts, Validity, and Minimum Competency Testing*. Ed. by G. F. Madaus. Boston, Massachussetts: Kluwer-Nijhoff. Chap. 10, pp. 197–208.
Chall, J. S., S. S. Conard, and S. H. Harris (1977). *An Analysis of Textbooks in Relation to Declining SAT Scores*. Tech. rep. Prepared fo the Advisory Panel on on the Scholastic Aptitude Test Score Decline. Princeton, New Jersey.
Chall, J. S. and E. Dale (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, Massachusetts: 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 (2018). "Patterns in Economics Journal Acceptances and Rejections". *American Economist* 63 (1), pp. 94–109.
Coate, S. and G. C. Loury (1993). "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review* 83 (5), pp. 1220–1240.
Coleman, E. B. (1964). "The Comprehensibility of Several Grammatical Transformations". *Journal of Applied Psychology* 48 (3), pp. 186–190.
— (1965). "Learning of Prose Written in Four Grammatical Transformations". *Journal of Applied Psychology* 49 (5), pp. 332–341.
Correll, S. and C. Simard (2016). "Vague Feedback Is Holding Women Back". *Harvard Business Review*. https://hbr.org/2016/04/research-vague-feedback-is-holding-women-back. Accessed: 2016-10-04.
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.
Cox, C. (2007). *Closing Remarks to the Second Annual Corporate Governance Summit*. Delivered at USC Marshall School of Business, Los Angeles, California, 23 March.
Dale, E. and J. S. Chall (1948). "A Formula for Predicting Readability". *Educational Research Bulletin* 27 (1), pp. 11–20.
De Franco, G. et al. (2015). "Analyst Report Readability". *Contemporary Accounting Research* 32 (1), pp. 76–104.
DuBay, W. H. (2004). *The Principles of Readability*. Costa Mesa, California: Impact Information.
Ductor, L., S. Goyal, and A. Prummer (2018). "Gender and Collaboration". Mimeo. Cambridge.
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.
Enke, B. (2018). "Moral Values and Voting: Trump and Beyond". NBER Working Paper Series, No. 24268.
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.
Fass, W. and G. M. Schumacher (1978). "Effects of Motivation, Subject Activity, and Readability on the Retention of Prose Materials". *Journal of Educational Psychology* 70 (5), pp. 803–807.
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.
Flesch, R. (1949). *The Art of Readable Writing*. New York, New York: Harper and Brothers Publishers.
Foschi, M. (1996). "Double Standards in the Evaluation of Men and Women". *Social Psychology Quarterly* 59 (3), pp. 237–254.
Foster, D. R. and D. H. Rhoney (2002). "Readability of Printed Patient Information for Epileptic Patients". *Annals of Pharmacotherapy* 36 (12), pp. 1856–1861.
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.
Funk, P. and H. Perrone (2017). "Gender Differences in Academic Performance : The Role of Personality Traits". Mimeo.
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. (2014a). "A Grand Gender Convergence: Its Last Chapter". *American Economic Review* 104 (4), pp. 1091–1119.
— (2014b). "A Pollution Theory of Discrimination: Male and Female Differences in Occupations and Earnings". In: *Human Capital in History: The American Record*. Ed. by L. P. Boustan, C. Frydman, and R. A. Margo. Cambridge, Massachusetts: National Bureau of Economic Research, pp. 313–348.
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.
Gorelkina, O. and E. Hengel (2018). "A reply to Tol, 2018". Mimeo.
Grossbard, S., T. Yilmazer, and L. Zhang (2018). "The Gender Gap in Citations of Economics Articles: Lessons from Economics of the Household". Mimeo.
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.
Guerini, M., A. Pepe, and B. Lepri (2012). "Do Linguistic Style and Readability of Scientific Abstracts Affect their Virality?" In: *Proceedings of the Sixth Interntaional AAAI Conference of Weblogs and Social Media*. Dublin, pp. 475–478.
Hamermesh, D. S. (1994). "Facts and Myths about Refereeing". *Journal of Economic Perspectives* 8 (1), pp. 153–163.
Hart, R. L. (2000). "Co-authorship in the Academic Library Literature: A Survey of Attitudes and Behaviors". *Journal of Academic Librarianship* 26 (5), pp. 339–345.
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). "Gender Differences in Peer Review Scrutiny". Mimeo.
— (2017). "Publishing while Female: Are Women Held to Higher Standards? Evidence From Peer Review." Mimeo.

— (2018). "Gender Differences in Citations at Top Economics Journals". Mimeo.
Hopewell, S. et al. (2014). "Impact of Peer Review on Reports of Randomised Trials Published in Open Peer Review Journals: Retrospective Before and After Study". *British Medical Journal* 349, g4145–g4145.
Hussin, M. F. et al. (2012). "The Readability of Transmission Line Characteristics Lab Manual". In: *IEEE Control and System Graduate Research Colloquium*. Shah Alam, Selangor, pp. 398–401.
Ittonen, K., E. Vähämaa, and S. Vähämaa (2013). "Female Auditors and Accruals Quality". *Accounting Horizons* 27 (2), pp. 205–228.
Jansen, D. J. (2011). "Does the Clarity of Central Bank Communication Affect Volatility in Financial Markets? Evidence from Humphrey-Hawkins Testimonies". *Contemporary Economic Policy* 29 (4), pp. 494–509.
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). *Editorial Report: 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.
Kemper, S. (1983). "Measuring the Inference Load of a Text". *Journal of Educational Psychology* 75 (3), pp. 391–401.
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.
Kintsch, W. and J. R. Miller (1984). "Readability: A View from Cognitive Psychology". In: *Understanding Reading Comprehension*. Ed. by J. Flood. Newark, Delaware: International Reading Association, pp. 220–232.
Klare, G. R. (1976). "Judging Readability". *Instructional Science* 5 (1), pp. 55–61.
Klare, G. R. and K. L. Smart (1973). "Analysis of the Readability Level of Selected USAFI Instructional Materials". *Journal of Educational Research* 67 (4), p. 176.
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.
Kumar, S. and K. Ratnavelu (2016). "Perceptions of Scholars in the Field of Economics on Coauthorship Associations: Evidence from an International Survey". *PLoS ONE* 11 (6), pp. 1–18.
Laband, D. N. and C. N. Taylor (1992). "The Impact of Bad Writing in Economics". *Economic Inquiry* 30 (4), pp. 673–688.
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.

Law, D. S. and D. Zaring (2010). "Law versus Ideology: the Supreme Court and the Use of Legislative History". *William and Mary Law Review* 51 (5), pp. 1653–1747.
Lawrence, A. (2013). "Individual Investors and Financial Disclosure". *Journal of Accounting and Economics* 56 (1), pp. 130–147.
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.
Lei, L. and S. Yan (2016). "Readability and Citations in Information Science: Evidence from Abstracts and Articles of Four Journals (2003–2012)". *Scientometrics* 108 (3), pp. 1155–1169.
Leuven, E. and B. Sianesi (2003). *PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing*. http://ideas.repec.org/c/boc/bocode/s432001.html. Accessed: 2018-03-20.
Li, F. (2008). "Annual Report Readability, Current Earnings, and Earnings Persistence". *Journal of Accounting and Economics* 45 (2-3), pp. 221–247.
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 (2016). "Textual Analysis in Accounting and Finance: A Survey". *Journal of Accounting Research* 54 (4), pp. 1187–1230.
Lundberg, S. J. (1991). "The Enforcement of Equal Opportunity Laws Under Imperfect Information: Affirmative Action and Alternatives". *Quarterly Journal of Economics* 106 (1), pp. 309–326.
Lundberg, S. J. (2017). "Committee on the Status of Women in the Economics Profession (CSWEP)". *American Economic Review* 107 (5), pp. 759–776.
Lundberg, S. J. and R. Startz (1983). "Private Discrimination and Social Intervention in Competitive Labor Markets". *American Economic Review* 73 (3), pp. 340–347.
McFadden, K. L. (1996). "Comparing Pilot-error Accident Rates of Male and Female Airline Pilots". *Omega* 24 (4), pp. 443–450.
Meade, C. D. and J. C. Byrd (1989). "Patient Literacy and the Readability of Smoking Education Literature". *American Journal of Public Health* 79 (2), pp. 204–206.
Meyer, B. J. F. (1982). "Reading Research and the Composition Teacher: The Importance of Plans". *College Composition and Communication* 33 (1), pp. 37–49.
Miller, B. P. (2010). "The Effects of Reporting Complexity on Small and Large Investor Trading". *Accounting Review* 85 (6), pp. 2107–2143.
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.
Plavén-Sigray, P. et al. (2017). "The Readability of Scientific Texts is Decreasing over Time". *eLife* 6 (e27725), pp. 1–14.
Püttmann, L. (2017). *VoxEU Gobbledygook*. http://lukaspuettmann.com/2017/12/09/voxeu-gobbledygook/. Accessed: 2018-05-19.
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.
Richardson, J. V. (1977). "Readability and Readership of Journals in Library Science". *Journal of Academic Librianship* 3 (1), pp. 20–22.
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.
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 (10), pp. 811–825.
Sarsons, H. (2017). "Gender Differences in Recognition for Group Work". Mimeo.
Sawyer, A. G., J. Laran, and J. Xu (2008). "The Readability of Marketing Journals: Are Award-Winning Articles Better Written?" *Journal of Marketing* 72 (1), pp. 108–117.
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.
Spirling, A. (2016). "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915". *Journal of Politics* 78 (1), pp. 120–136.
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.
Swanson, C. E. (1948). "Readability and Readership: A Controlled Experiment". *Journalism Bulletin* 25 (4), pp. 339–343.

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.
Tol, R. S. (2018). "Gender at Energy Economics". *Energy Economics* 72, pp. 558–559.
Torgler, B. and M. Piatti (2013). *A Century of American Economic Review*. New York, New York: Palgrave Macmillan.
Tsugawa, Y. et al. (2017). "Comparison of Hospital Mortality and Readmission Rates for Medicare Patients Treated by Male vs Female Physicians". *JAMA Internal Medicine* 177 (2), pp. 206–213.
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.
Velte, P. (2018). "Does Gender Diversity in the Audit Committee Influence Key Audit Matters' Readability in the Audit Report? UK Evidence". *Corporate Social Responsibility and Environmental Management* (forthcoming).
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.
Wallace, L. S. et al. (2008). "Suitability and Readability of Consumer Medical Information Accompanying Prescription Medication Samples". *Patient Education and Counseling* 70 (3), pp. 420–425.
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.
Weisberg, Y. J., C. G. De Young, and J. B. Hirsh (2011). "Gender Differences in Personality across the Ten Aspects of the Big Five". *Frontiers in Psychology* 2 (178).
Weisshaar, K. (2017). "Publish and Perish? An Assessment of Gender Gaps in Promotion to Tenure in Academia". *Social Forces* 96 (2), pp. 529–560.
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, Massachusetts: Harvard University Press.
Zipf, G. Z. (1935). *The Psycho-Biology of Language: an Introduction to Dynamic Philology*. Boston, Massachusetts: MIT Press.

Appendices

A Research assessing readability score validity

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

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

Figure 2 should convincingly convey the message that the readability scores I use in this paper positively correlate with other—arguably more reliable—measures of text difficulty. Athough I have made every effort to include all relevant research (subject to the restrictions above), I do not promise that the studies listed below represent a completely comprehensive review.

Studies included in Figure 2

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

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

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

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

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

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

Chall, J. S. and E. Dale (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, Massachusetts: Brookline Books.

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

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

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

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

Cunningham, J. W., E. H. Hiebert, and H. A. Mesmer (2018). "Investigating the Validity of Two Widely Used Quantitative Text Tools". *Reading and Writing* 31 (4), pp. 813–833.
Dale, E. and J. S. Chall (1948). "A Formula for Predicting Readability". *Educational Research Bulletin* 27 (1), pp. 11–20.
Dale, E. and R. W. Tyler (1934). "A Study of the Factors Influencing the Difficulty of Reading Materials for Adults of Limited Reading Ability". *Library Quarterly: Information, Community, Policy* 4 (3), pp. 384–412.
Entin, E. B. and G. R. Klare (1978). "Some Inter-relationships of Readability, Cloze and Multiple Choice Scores on a Reading Comprehension Test". *Journal of Literacy Research* 10 (4), pp. 417–436.
Flesch, R. (1948). "A New Readability Yardstick". *Journal of Applied Psychology* 32 (3), pp. 221–233.
Froese, V. (1971). *Cloze Readability versus the Dale-Chall Formula*. Tech. rep. Winnipeg, Manitoba: University of Manitoba.
Fulcher, G. (1997). "Text Difficulty and Accessibility: Reading Formulae and Expert Judgement". *System* 25 (4), pp. 497–513.
Gray, W. W. and B. E. Leary (1935). *What Makes a Book Readable*. Chicago, Illinois: University of Chicago Press.
Greenfield, J. (1999). "Classic Readability Formulas in an EFL Context: Are They Valid for Japanese Speakers?" PhD thesis. Temple University.
— (2004). "Readability Formulas for EFL". *JALT Journal* 26 (1), pp. 5–24.
Guthrie, J. T. (1972). "Learnability versus Readability of Texts". *Journal of Educational Research* 65 (6), pp. 273–280.
Harris, A. J. and M. D. Jacobson (1976). "Predicting Twelfth Graders' Comprehension Scores". *Journal of Reading* 20 (1), pp. 43–46.
Harwell, M. R. et al. (1996). "Evaluating Statistics Texts Used in Education". *Journal of Educational and Behavioral Statistics* 21 (1), pp. 3–34.
Hayes, D. P., L. T. Wolfer, and M. F. Wolfe (1996). "Schoolbook Simplification and Its Relation to the Decline in SAT-Verbal Scores". *American Educational Research Journal* 33 (2), pp. 489–508.
Hengel, E. (2017). "Publishing while Female: Are Women Held to Higher Standards? Evidence From Peer Review." Mimeo.
Hull, L. C. (1979). "Beyond Readability: Measuring the Difficulty of Technical Writing". PhD thesis. Renesselaer Polytechnic Institute.
Janan, D. and D. Wray (2014). "Reassessing the Accuracy and Use of Readability Formulae". *Malaysian Journal of Learning and Instruction* 11 (1), pp. 127–145.
Jongsma, E. A. (1972). "The Difficulty of Children's Books: Librarians' Judgments vs. Formula Estimates". *Elementary English* 49 (1), pp. 20–26.
Kanouse, D. E. et al. (1981). *Informing Patients about Drugs: Summary Report on Alternative Designs for Prescription Drug Leaflets*. Tech. rep. Santa Monica, California: Rand Corporation.
Kemper, S. (1983). "Measuring the Inference Load of a Text". *Journal of Educational Psychology* 75 (3), pp. 391–401.
Kincaid, J. P. et al. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Tech. rep. Memphis, Tennessee: Naval Technical Training Command.
King, M. M., A. S. W. Winton, and A. D. Adkins (2003). "Assessing the Readability of Mental Health Internet Brochures for Children and Adolescents". *Journal of Child and Family Studies* 12 (1), pp. 91–99.

Klare, G. R. (1952). "Measures of the Readability of Written Communication: An Evaluation". *Journal of Educational Psychology* 43 (7), pp. 385–399.

Klingbeil, C., M. W. Speece, and H. Schubiner (1995). "Readability of Pediatric Patient Education Materials. Current Perspectives on an Old Problem." *Clinical Pediatrics* 34 (2), pp. 96–102.

Lee, W. D. and B. R. Belden (1966). "A Cross-Validation Readability Study of General Psychology Textbook Material and the Dale-Chall Readability Formula". *Journal of Educational Research* 59 (8), pp. 369–373.

Lenzner, T. (2014). "Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty?" *Sociological Methods and Research* 43 (4), pp. 677–698.

Ley, P. and T. Florio (1996). "The Use of Readability Rormulas in Health Care". *Psychology, Health and Medicine* 1 (1), pp. 7–28.

Lorge, I. (1948). "The Lorge and Flesch Readability Formulas: A Correction". *School & Society* 67, pp. 141–142.

McLaughlin, G. (1969). "SMOG Grading: A New Readability Formula". *Journal of Reading* 12 (8), pp. 639–646.

Meade, C. D. and J. C. Byrd (1989). "Patient Literacy and the Readability of Smoking Education Literature". *American Journal of Public Health* 79 (2), pp. 204–206.

Meade, C. D. and C. F. Smith (1991). "Readability Formulas: Cautions and Criteria". *Patient Education and Counseling* 17 (2), pp. 153–158.

Miller, L. R. (1974). "Predictive Powers of the Flesch and Bormuth Readability Formulas". *International Journal of Business Communication* 11 (2), pp. 21–30.

Morris, L. A., A. Myers, and D. G. Thilman (1980). "Application of the Readability Concept to Patient-Oriented Drug Information". *American Journal of Health-System Pharmacy* 37 (11), pp. 1504–1509.

Powell-Smith, K. A. and K. L. Bradley-Klug (2001). "Another Look at the 'C' in CBM: Does It Really Matter if Curriculum-based Measurement Reading Probes Are Curriculum-based?" *Psychology in the Schools* 38 (4), pp. 299–312.

Powers, R. D., W. A. Sumner, and B. E. Kearl (1958). "A Recalculation of Four Readability Formulas". *Journal of Educational Psychology* 49 (2), pp. 99–105.

Russell, D. H. and H. R. Fea (1951). "Validity of Six Readability Formulas as Measures of Juvenile Fiction". *Elementary School Journal* 52 (3), pp. 136–144.

Singer, H. (1975). "The Seer Technique: A Non-Computational Procedure for Quickly Estimating Readability Level". *Journal of Reading Behavior* 7 (3), pp. 255–267.

Singh, J. (2003). "Reading Grade Level and Readability of Printed Cancer Education Materials". *Oncology Nursing Forum* 30 (5), pp. 867–870.

Štajner, S. et al. (2012). "What Can Readability Measures Really Tell Us About Text Complexity?" In: *Workshop on Natural Language Processing for Improving Textual Accessibility*, pp. 14–21.

Sullivan, R. J. (1976). "A Comparison of Results Obtained Using the Cloze Procedure with Readability Levels Using the Dale-Chall Formula on Selected University Textbooks". In: *26th Annual Meeting of the National Reading Conference*. Atlanta, Georgia.

Van Oosten, P., D. Tanghe, and V. Hoste (2010). "Towards an Improved Methodology for Automated Readability Prediction". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pp. 775–782.

Wells, J. A. (1994). "Readability of HIV/AIDS Educational Materials: The Role of the Medium of Communication, Target Audience, and Producer Characteristics". *Patient Education and Counseling* 24 (3), pp. 249–259.

Woods, B., G. Moscardo, and T. Greenwood (1998). "A Critical Review of Readability and Comprehensibility Tests". *Journal of Tourism Studies* 9 (2), pp. 49–61.
Zheng, J. and H. Yu (2017). "Readability Formulas and User Perceptions of Electronic Health Records Difficulty: A Corpus Study". *Journal of Medical Internet Research* 19 (3), pp. 1–15.

B Calculating readability scores

As discussed in Section 2.1, I wrote the Python module `Textatistic` to transparently calculate the readability scores in this study. The code and documentation are available on GitHub; I provide a brief description here.

To determine sentence count, the program replaces common abbreviations with their full text,[1] decimals with a zero and deletes question and exclamation marks used in an obvious, mid-sentence rhetorical manner.[2] 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.[3]

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.

[1] 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.

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

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

C Description of control variables

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

In total, 1,039 different institutions were identified. For each institution, I count the number of articles in which it was listed as an affiliation in a given year and smooth the average over a five-year period.[4] Institutions are ranked on an annual basis using this figure and then grouped to create fifteen dynamic dummy variables. Institutions ranked in positions 1–9 are assigned individual dummy variables. Those in positions 10–59 are grouped in bins of 10 to form six dummy variables. Institutions ranked 60 or above were collectively grouped to form a final dummy variable.[5] 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).[6] In Section 4.3 and Section 4.6, 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.[7]

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;[8] 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).[9]

I create dummy variables corresponding to the 20 primary and over 700 tertiary *JEL* categories to control for subject matter. The *JEL* system was significantly revised in 1990; because

[4] 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.

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

[6] 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).

[7] 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.

[8] 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).

[9] 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.

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.[10] 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.

To control for motherhood's impact on revision times, I recorded children's birth years for women with at least one entirely female-authored paper in *Econometrica*. I personally (and, I apologise, rather unsettlingly) gleaned this information from published profiles, CVs, acknowledgements, Wikipedia, personal websites, Facebook pages, background checks and local school district/popular extra-curricular activity websites.[11] 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.[12]

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

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

[12] 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 most of the other data in this paper).

D Proofs for Theorem 1 and Corollary 1

The proof of Theorem 1 follows directly from Lemma 5, at the end of this section. The proof of Lemma 5 relies on a series of additional lemmas stated and proved below. It is followed by a proof of Corollary 1. Throughout, $\{(r_{0it}, R_{it})\}$ represents the sequence of readability choices made by author i for all t. $R_i^\star$ 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 (10). It satisfies the following first order condition

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

where v_{1it}^s represents Equation (10) 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,

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

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

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

v_{1it}^s is non-negative;[13] 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 (D.3) is non-negative. Since $c_i'(r) < \phi_i'(r)$ for all $r < R_i^\star$, the left-hand side of Equation (D.1) is strictly positive for all $r < R_i^\star$, so r_{0it} must be at least as large as $R_i^\star$.

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

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

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

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

Similarly, because $\Pi_{0it}^s(r_{0it}) \Pi_{1it}^s(R_{it}) \leq 1$ for all s and $\phi_i(r)$ and $c_i(r)$ are finite at all $r < \infty$, Equation (10) is likewise finite for all $r < \infty$. Thus,

$$\sup \left\{ \operatorname*{argmax}_{r_{0it}} \int_\Sigma \Pi_{0it}^s(r_{0it}) v_{1it}^s \, \mathrm{d}\mu_i + \phi_i(r_{0it}) - c_i(r_{0it}) \right\} < \infty,$$

[13] Equation (9) evaluated at $r_{1it} = 0$ is non-negative. Since r_{1it} maximises Equation (9), v_{1it}^s is likewise non-negative.

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 (9) is less than 0 for all $r_{1it} > 0$. Since $R_i^\star \leq r_{0it}$ and $\Pi_{1it}^s(R_{it}) \leq 1$

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

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

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

Lemma 2. *$r_{0i} \leq r_{0it}$ and $R_i^s \leq R_{it}^s$ 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 $\{(r_{0it}, R_{it}^{q^\star})\}$ denote the complete subsequence of $\{(r_{0it}, R_{it})\}$ in which state q is reached. Thus,

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

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

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

$$\pi_{1it}^s(R_{it}^s) = \frac{1}{u_i}\left(c_i'(R_{it}^s) - \phi_i'(R_{it}^s)\right). \tag{D.7}$$

Let π_{1i}^s denote the terminal value of π_{1it}^s as t tends to ∞. π_{1i}^s is finite; thus, $\{\pi_{1it}^s\}$ itself converges: if $\widetilde{R}_i^s < R_i^s$, then $\pi_{1it}^s(R_{it}^s) = 0$ for all $t > t''$, where t'' has been redefined to assure $\widetilde{R}_i^s \leq R_{it}^s$; if $R_i^s \leq \widetilde{R}_i^s$ and $\pi_{1i}^s(R_i^s) = \infty$, then $\pi_{1i}^s(R) = 0$ for all $R > R_i^s$, a contradiction (see Footnote 77).

Convergence by $\{\pi_{1it}^s\}$ and $\{R_{it}^s\}$ means

$$\lim_{t\to\infty} \left|\pi_{1it+1}^s(R_{it+1}^s) - \pi_{it}^s(R_{it}^s)\right| = 0.$$

Yet Equation (D.7) implies

$$\begin{aligned} &\lim_{t\to\infty} \left|\pi_{1it+1}^s(R_{it+1}^s) - \pi_{it}^s(R_{it}^s)\right| \\ &\quad = \lim_{\varepsilon\to 0} \frac{1}{u_i}\Big(\left[c_i'(R_{it}^s + \varepsilon) - c_i'(R_{it}^s)\right] - \left[\phi_i'(R_{it}^s + \varepsilon) - \phi_i'(R_{it}^s)\right]\Big) \\ &\quad = \frac{1}{u_i}\left(c_i''(R_i^s) - \phi_i''(R_i^s)\right), \end{aligned} \tag{D.8}$$

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

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

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

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

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 (D.2), Equation (D.9) is equivalent to

$$K + \mu^s_i \pi^s_{0it}(r_{0it}) v^s_{1it} = \Big(1 - \mu^s_i \Pi^s_{0it}(r_{0it}) \Big) \Big(c'_i(r_{0it}) - \phi'_i(r_{0it}) \Big). \tag{D.10}$$

If $r^s_{1it} = 0$ then $r^s_{1it+1} = 0$, and $\partial v^s_{1it} / \partial r_{1it} = \pi^s_{1it}(R^s_{it}) u_i$.[14] In this case, Equation (D.9) is equivalent to

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

By the monotone convergence theorem, $\{v^s_{1it}\}$ and $\{\Pi^s_{0it}\}$ converge.[15] If $\widetilde{r}^s_{0i} < r_{0i}$, then $\pi^s_{0it}(r_{0it}) = 0$ for all $t > t'$, where t' has been redefined to assure $\widetilde{r}^s_{0i} \leq r_{0it}$; if $r_{0i} \leq \widetilde{r}^s_{0i}$, then

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

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

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

and

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

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

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

For the moment, assume there exists t'' such that for all $r \in (r_{0it''}, r_{0i})$, K is constant.[16] Thus, changes over time to the left-hand sides of Equation (D.10) and Equation (D.11) converge to 0. Yet the right-hand sides of Equation (D.10) and Equation (D.11) do not, since

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

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

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

is strictly greater than 0, where convergence of $\{r_{0it}\}$ guarantees that for all (sufficiently small) $\varepsilon > 0$ there exists $r_{0it+1} = r_{0it} + \varepsilon$.[17] 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 (D.10) and Equation (D.11) converge to 0 while changes to their right-hand sides do not. Because the number of states is finite by assumption, this establishes the general contradiction. □

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

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

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

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

Proof. Suppose $\{(r_{0it}, R_{it}^{s^\star})\}$ has two cluster points: $(r_{0i}', R_i^{s\prime})$ 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 (9) and Equation (10) 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.[18] Since the choice of s was arbitrary exists a unique cluster point of $\{(r_{0it}, R_{it}^{s^\star})\}$ for every $s^\star \in \Sigma$. □

[16] 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.

[17] 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 (D.12) shows is not possible.

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

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

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

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

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

and

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

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

$$\lim_{\varepsilon \to 0-} \Pi^q_{1i}(R^q_i + \varepsilon) < 1.^{19} \tag{D.15}$$

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

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

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 76), Equation (D.16) necessarily implies

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

a contradiction.

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

Proof of Corollary 1. I first show that Equation (12) conservatively estimates D_{ik} when $\Sigma_{A_{it}} \subset \Sigma_{A_{kt}}$. Let $r_{0it} < R_{it}$. From Equation (11) and the definition of δ^s_{1ik},

$$\begin{aligned} R_{it} - R_{kt} &= \widetilde{R}^s_i + e_{1it} - \max\left\{R^\star_k, \widetilde{r}^{\bar{s}_k}_{0k} + e_{0kt}, \widetilde{R}^s_k + e_{1kt}\right\} \\ &\leq \widetilde{R}^s_i - \widetilde{R}^s_k + e_{1it} - e_{1kt} \\ &= \delta^s_{1ik} + e_{1it} - e_{1kt}. \end{aligned} \tag{D.17}$$

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

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

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

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 (D.18) is equivalent to

$$\begin{aligned} R_{it} - R_{kt} &\leq \widetilde{r}^{\overline{s}_i}_{0i} - \widetilde{r}^{\overline{s}_k}_{0k} + e_{0it} - e_{0kt} \\ &= \delta^{\overline{s}_i}_{0ik} + \widetilde{r}^{\overline{s}_i}_{0k} - \widetilde{r}^{\overline{s}_k}_{0k} + e_{0it} - e_{0kt}. \end{aligned} \tag{D.19}$$

$e_{0it} = e_{0kt}$ 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 (D.19) implies $R_{it} - R_{kt} \leq \delta^{\overline{s}_i}_{0ik}$ if $R_{it} = r_{0it}$. Meanwhile, Equation (D.17) 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.

□

E Gender and readability, by *JEL* code

Figure E.1 displays results from an ordinary least squares regression on the Dale-Chall score; regressors are: (i) ratio of female co-authors; (ii) dummies for each primary *JEL* code; (iii) interactions from (i) and (ii); (iv) controls for editor, journal, year, institution and English fluency; and (v) quality controls—citation count and max. T_j fixed effects.[20] Due to small samples—particularly of female authors—Figure E.1 includes 561 articles from *AER Papers & Proceedings*.[21]

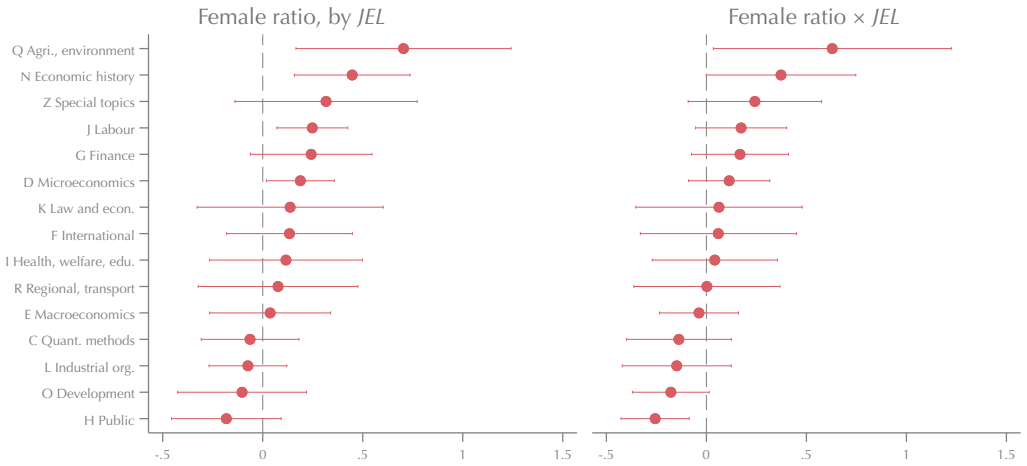

FIGURE E.1: Gender differences in readability, by *JEL* classification

Notes. Sample 5,777 articles, including 561 from *AER Papers & Proceedings* (see Footnote 53). Codes A, B, M and P dropped due to small sample sizes of female-authored papers (see Footnote 20). 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 (asinh) 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 mean effect at observed *JEL* codes is 0.14 (standard error 0.048). Right-hand graph displays interaction terms (β_3^k). Horizontal lines represent 90 percent confidence intervals from standard errors adjusted for clustering on editor.

Points reflect marginal effects across *JEL* classification; bars represent 90 percent confidence intervals from standard errors clustered by editor. The mean effect at observed *JEL* codes is 0.14 (standard error 0.048). This estimate coincides with results in Table 4—women's papers require six fewer weeks of schooling to understand—and is highly significant.

Women earn higher marks for clarity in 11 out of 15 categories; only three are at least 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 E.1'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 E.1 suggests two things. First, the mostly insignificant interaction terms indicate outlier fields are probably not driving journals' gender readability gap—nor is any specific

[20] 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.

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

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.

F Double- and semi-blind review

In an earlier version of this paper (Hengel, 2015), I compared readability gaps in published articles subjected to blind review—*i.e.*, double-blind review before the internet—and non-blind review—*i.e.*, single-blind review or double-blind review after the internet. Blind review appeared to exacerbate the gender readability gap.

These findings, however, were not robust to including fixed effects for year of publication. Table F.1 repeats the analysis from Hengel (2015, Table 3.9 (first panel), p. 65) including them. Figures represent the marginal effect of female ratio in non-blind (β_{1P}) and blind ($\beta_{1P} + \beta_{3P}$) review from OLS estimation of Equation (F.1):

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

The gender readability gap is positive and significant in non-blind review; when papers are blindly evaluated, however, the gap is generally smaller and never significant. Difference-in-difference estimates (Table F.1, third row) are mostly positive, indicating—in contrast to results in Hengel (2015) but consistent with results in Table 7—that the readability gap was smaller under blinded peer review. Nevertheless, standard errors are large relative to the size of the effect; please interpret these results with caution.

Table F.1: The impact of double-blind review in all published articles

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Non-blind	0.85	0.15	0.28*	0.19*	0.09*
	(0.57)	(0.13)	(0.15)	(0.11)	(0.05)
Blind	0.49	0.03	0.18	0.03	0.19**
	(1.03)	(0.24)	(0.32)	(0.22)	(0.08)
Difference	0.36	0.12	0.10	0.16	−0.10
	(1.13)	(0.27)	(0.35)	(0.24)	(0.09)
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Quality controls	✓[3]	✓[3]	✓[3]	✓[3]	✓[3]
Native speaker	✓	✓	✓	✓	✓

Notes. Sample 9,122 articles. Columns display the marginal effect on female ratio for papers undergoing non-blind review (β_{1P}) and blind review ($\beta_{1P} + \beta_{3P}$) from OLS estimation of Equation (F.1). Quality controls denoted by ✓[3] include max. t_j, only. Standard errors clustered on year in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Table 7 and Table F.1 suggest double-blind review may have successfully reduced peer review's impact on the gender readability gap *before* the internet. Unfortunately, it has been less effective *after* the internet. I dropped NBER–published article pairs published pre-internet (*i.e.*, before Google's incorporation in 1998) and replicated Table 7 with Blind_j equal to 1 if article j was subjected to an official policy of double-blind review after the internet. The results, presented in Table F.2, suggests a positive gender readability gap in both samples. If anything, blinded peer review coupled with an easy alternative for determining authors' identities seems to exacerbate gender differences.

TABLE F.2: The impact of double-blind review after the internet

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Non-blind	0.76	0.29	0.24	0.08	0.15**
	(0.92)	(0.29)	(0.31)	(0.18)	(0.07)
Blind post-internet	1.11	0.59**	0.58*	0.40*	0.01
	(1.11)	(0.29)	(0.34)	(0.23)	(0.13)
Difference	−0.35	−0.30	−0.34	−0.32	0.14
	(1.55)	(0.43)	(0.50)	(0.32)	(0.17)
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Quality controls	✓[3]	✓[3]	✓[3]	✓[3]	✓[3]
Native speaker	✓	✓	✓	✓	✓

Notes. Sample 1,380 NBER working papers; 1,378 published articles. Table replicates Table 7 with Blind_j equal to 1 if article j was subjected to an official policy of double-blind review after the internet. (NBER–published article pairs published pre-internet are dropped.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

G Abstract word limits

I attribute the change in readability between draft and final versions of a paper to the peer review process.[22] Yet NBER working paper abstracts can be of any length while abstracts published in *Econometrica* and *AER* cannot—they are restricted to 150 and 100 words, respectively. Observed readability gaps could consequently result from gender differences in how authors conform to these limits.

To test this hypothesis, I replicated the analysis described Section 4.3.3 (and shown in Table 6) on the subset of articles with draft abstracts below the official minimum word limit of the journals in which they were eventually published. Results are shown in Table G.1. Despite dropping about 40 percent of observations, coefficient magnitudes are similar to those reported in Table 6; standard errors are somewhat larger.

Table G.1: Table 6, draft abstracts below official word limits

	OLS	FGLS			OLS
	Published article	Working paper	Published article	Difference	Change in score
Flesch Reading Ease	0.89	2.31	2.84*	0.52	0.54
	(0.88)	(1.50)	(1.56)	(0.83)	(0.86)
Flesch-Kincaid	0.55**	0.06	0.59*	0.53**	0.54*
	(0.27)	(0.35)	(0.33)	(0.27)	(0.28)
Gunning Fog	0.57**	0.21	0.73**	0.52**	0.52*
	(0.24)	(0.39)	(0.34)	(0.26)	(0.27)
SMOG	0.28*	0.22	0.45**	0.23	0.24
	(0.15)	(0.27)	(0.22)	(0.16)	(0.16)
Dale-Chall	0.23***	0.34***	0.51***	0.17**	0.17**
	(0.08)	(0.12)	(0.12)	(0.07)	(0.08)
Editor effects	✓	✓	✓		✓
Journal effects	✓	✓	✓		✓
Year effects	✓	✓	✓		
Journal×Year effects	✓	✓	✓		✓
Quality controls	✓[2]	✓[2]	✓[2]		✓[3]
Native speaker	✓	✓	✓		✓

Notes. Sample 1,067 NBER working papers; 1,065 published articles. Estimates are identical to those in Table 6, except that the sample includes only papers with an NBER abstract below the official minimum word limit of the journal in which it was eventually published. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

[22] See Section 4.3.4 for a more detailed discussion and justification of the assumptions underpinning this claim.

H Evaluating Theorem 1 using the entire sample of authors

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.

TABLE H.1: 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
	(1.15)	(0.24)	(0.29)	(0.21)	(0.10)
Men	39.37	−13.77	−17.54	−15.35	−10.99
	(0.31)	(0.07)	(0.08)	(0.06)	(0.03)
Average mean score					
Women	41.20	−13.36	−16.92	−14.92	−10.91
	(0.72)	(0.15)	(0.19)	(0.14)	(0.07)
Men	39.59	−13.69	−17.41	−15.26	−11.01
	(0.19)	(0.04)	(0.05)	(0.03)	(0.02)
Average final paper score					
Women	41.99	−13.10	−16.58	−14.66	−10.90
	(1.06)	(0.21)	(0.25)	(0.18)	(0.11)
Men	39.54	−13.71	−17.40	−15.24	−11.08
	(0.33)	(0.08)	(0.09)	(0.06)	(0.03)

Notes. Sample 1,675 authors; includes only authors with three or more publications. Figures are average readability scores for authors' first, mean and last published articles. Grade-level scores have been multiplied by negative one (see Section 2). Standard errors in parentheses.

Table H.1 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). Sample excludes authors with fewer than three publications.

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

Figure 7 plots mean Flesch Reading Ease scores grouped by authors' tth article; as the count increases, men and women diverge.[23] Table H.2 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 4.2), 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. Between columns ($t = 1$) and ($t = 2$), the gap marginally widens but

[23] 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 4.2. The results and conclusions were similar to those presented here.

TABLE H.2: Gender gap in readability at increasing t

	$t=1$	$t=2$	$t=3$	$t=4$–5	$t\geq 6$	All
Flesch Reading Ease	0.52	1.89**	5.04***	2.63	3.55	1.78**
	(0.67)	(0.85)	(1.21)	(1.95)	(2.21)	(0.73)
Flesch-Kincaid	0.10	0.18	0.88***	0.55	0.65	0.21
	(0.15)	(0.21)	(0.22)	(0.40)	(0.40)	(0.15)
Gunning Fog	0.24	0.38	1.13***	0.82*	0.90*	0.43**
	(0.17)	(0.24)	(0.27)	(0.44)	(0.49)	(0.18)
SMOG	0.16	0.28*	0.74***	0.64*	0.67*	0.33**
	(0.12)	(0.17)	(0.19)	(0.35)	(0.35)	(0.13)
Dale-Chall	0.09	0.13*	0.42***	0.28*	0.45*	0.19***
	(0.06)	(0.08)	(0.14)	(0.16)	(0.23)	(0.07)
No. observations	6,877	2,828	1,675	1,908	2,777	12,016
N_j	✓	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓
Journal×Year effects						✓
Institution effects	✓	✓	✓	✓	✓	✓
Quality controls	✓[4]	✓[4]	✓[4]	✓[4]	✓[4]	✓[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 4.2). 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 (asinh) and max. T_j fixed effects; ✓[4] includes citation count (asinh), only. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

is not significant; after that, it triples (at least); the increase is significant ($p<0.05$) for all five scores (Table H.3).[24] 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.[25]

First-time publications are not driving the observed readability gap. Figure 7 suggests little or no gender difference when $t=1$; Table H.2 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 (Table H.3) 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.

[24] Figures in columns ($t=2$) and ($t=3$) of Table H.2 are roughly in line with third column estimates in Table 6—on average, $t=2.7$ for female-authored articles released first as NBER working papers.

[25] 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 H.3: Table H.3, equality test statistics

	$t = 1$ vs. 2	1 vs. 3	1 vs. 4–5	1 vs. ≥ 6	2 vs. 3
Flesch Reading Ease	2.094	9.931	1.176	1.792	5.391
Flesch-Kincaid	0.158	9.923	1.071	1.391	7.439
Gunning Fog	0.344	7.714	1.310	1.539	5.448
SMOG	0.597	6.875	1.544	2.051	4.280
Dale-Chall	0.163	4.893	1.396	2.293	3.889

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

I Alternative program for calculating readability scores

In this section, I replicate Table 3, Table 4, Table 6 and Table H.2 using readability scores generated by the R `readability` package, an alternative program for calculating Flesch-Kincaid, Gunning Fog and SMOG readability scores.[26] Replications for other tables and figures presented in the paper are not shown, but will be made available on request.

`Textatistic` and `readability` employ different strategies to adapt the scores to automated calculation—*e.g.*, `readability` counts semi-colons and dashes as sentence-ending terminations; `Textatistic` does not.[27] Results appear robust to these (and other) small discrepancies: coefficients are similar to those presented in the body of the paper; standard errors are usually smaller.

TABLE I.1: Table 3, alternative program for calculating readability

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Flesch-Kincaid	0.16	0.16	0.16	0.17	0.20	0.22	0.23*	0.25
	(0.11)	(0.11)	(0.11)	(0.11)	(0.12)	(0.14)	(0.13)	(0.15)
Gunning Fog	0.30**	0.30**	0.31**	0.33***	0.36***	0.33**	0.31**	0.29*
	(0.12)	(0.11)	(0.12)	(0.12)	(0.13)	(0.15)	(0.14)	(0.17)
SMOG	0.20**	0.20**	0.20**	0.21**	0.24**	0.21*	0.20*	0.19
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.11)	(0.10)	(0.13)
Editor effects	✓	✓	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓	✓	✓
Institution effects				✓	✓	✓	✓	✓
Quality controls					✓[1]	✓[1]	✓[1]	✓[1]
Native speaker					✓	✓	✓	✓
JEL (primary) effects							✓	
JEL (tertiary) effects								✓

Notes. 9,122 articles in (1)–(5); 5,216 articles in (6) and (7); 5,777 articles—including 561 from *AER Papers & Proceedings* (see Footnote 53)—in (8). Figures are identical to those in Table 3, except readability scores were calculated using the R `readability` program. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

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

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

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

	Flesch-Kincaid	Gunning Fog	SMOG
Female ratio (women)	0.40**	0.63**	0.40**
	(0.20)	(0.24)	(0.18)
Female ratio (men)	0.15	0.33	0.22
	(0.27)	(0.30)	(0.20)
Female ratio×male	−0.25	−0.30	−0.18
	(0.33)	(0.36)	(0.24)
Lagged score	0.05***	0.04**	0.04**
	(0.02)	(0.02)	(0.02)
z-test for no serial correlation			
Order 1	−15.79	−16.75	−18.70
Order 2	0.22	0.40	0.30
N_j	✓	✓	✓
Editor effects	✓	✓	✓
Journal effects	✓	✓	✓
Year effects	✓	✓	✓
Journal×Year effects	✓	✓	✓
Institution effects	✓	✓	✓
Quality controls	✓[1]	✓[1]	✓[1]
Native speaker	✓	✓	✓

Notes. Sample 9,188 observations (2,828 authors). Figures are identical to those in Table 4, except readability scores were calculated using the R `readability` program. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

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

	OLS	FGLS			OLS
	Published article	Working paper	Published article	Difference	Change in score
Flesch-Kincaid	0.52***	0.48**	0.89***	0.40**	0.40**
	(0.17)	(0.22)	(0.29)	(0.18)	(0.18)
Gunning Fog	0.52***	0.64***	1.02***	0.38**	0.37**
	(0.18)	(0.25)	(0.28)	(0.19)	(0.19)
SMOG	0.37***	0.42***	0.70***	0.28**	0.27**
	(0.14)	(0.16)	(0.19)	(0.13)	(0.13)
Editor effects	✓	✓	✓		✓
Journal effects	✓	✓	✓		✓
Year effects	✓	✓	✓		
Journal×Year effects	✓	✓	✓		✓
Quality controls	✓[2]	✓[2]	✓[2]		✓[3]
Native speaker	✓	✓	✓		✓

Notes. Sample 1,709 NBER working papers; 1,707 published articles. Figures are identical to those in Table 6, except readability scores were calculated using the R `readability` program. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

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

	$t = 1$	$t = 2$	$t = 3$	$t = 4$–5	$t \geq 6$	All
Flesch-Kincaid	0.11	0.22	0.83***	0.45	0.73*	0.20
	(0.15)	(0.23)	(0.26)	(0.42)	(0.39)	(0.16)
Gunning Fog	0.30*	0.36	1.07***	0.65	0.96*	0.43**
	(0.17)	(0.25)	(0.36)	(0.50)	(0.51)	(0.19)
SMOG	0.20*	0.24	0.67***	0.45	0.69*	0.29**
	(0.12)	(0.18)	(0.24)	(0.36)	(0.35)	(0.13)
No. observations	6,877	2,828	1,675	1,908	2,777	12,016
N_j	✓	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓
Journal×Year effects						✓
Institution effects	✓	✓	✓	✓	✓	✓
Quality controls	✓[4]	✓[4]	✓[4]	✓[4]	✓[4]	✓[1]
Native speaker	✓	✓	✓	✓	✓	✓

Notes. Figures are identical to those in Table H.2, except readability scores were calculated using the R `readability` program. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

J Alternative proxies for article gender

The following sections replicate Table 3, Table 4, Table 6, Table 10 and Table H.2 using alternative proxies for article gender. Additional replications not shown will be made available on request (subject to feasibility).

- In Appendix J.1, article gender is represented by a binary variable equal to one if at least half of all authors are female; mixed-gendered articles below this threshold are excluded.[28]
- In Appendix J.2, a paper is considered "female" if at least one author is female.
- In Appendix J.3, papers authored entirely by women are compared to papers authored entirely by men. Co-authored articles are excluded.[29]

The estimation strategy in Section 4.4 relies on within-author differences in readability scores at two specific t ($t = 1$ and $t = 3$). Because only a small number of women have majority and exclusively female-authored papers for both t, I reproduce instead results from Table H.2 on all three samples.[30]

In general, standard errors are smaller and coefficients larger in Appendix J.1 and Appendix J.2; the reverse is usually—but not always—true for Appendix J.3 (which includes a much smaller number of female-authored papers). The gender readability gap is consistently positive—with one exception: the re-estimation of Table 6's final two columns for the Flesch Reading Ease score in Appendix J.3.[31] Given the small number of female-authored manuscripts in this particular sample (54) combined with the large number of estimates shown in the paper, it's probably stranger there aren't any more.[32]

[28] Co-authored articles below this threshold are included—but not considered "female"—when estimating Table 4. This is done in order to generate within-author variation.

[29] Mixed-sex articles are included as "male" when estimating Table 4. (See Footnote 28.)

[30] See Footnote 103 for further discussion and results from a rough attempt to replicate the analysis in Section 4.4.2 on subsets of articles satisfying alternative gender conditions.

[31] Exclusively female-authored articles released as NBER working papers are also disproportionately authored by senior women—*i.e.*, the women for whom peer review's direct effect is smallest (Section 4.5).

[32] The coefficient on female-authorship for the Flesch Reading Ease score in the re-estimation of Table H.2's first column ($t = 1$) in Appendix J.3 is also negative. Although almost certainly due to chance, the result is nevertheless consistent with higher standards—*i.e.*, women learn about referees' standards over time; smaller total readability gaps should therefore be observed in their earliest papers.

J.1 Majority female-authored

TABLE J.1: Table 3, majority female-authored

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Flesch Reading Ease	0.91***	0.90***	0.87**	0.90***	1.01***	0.74*	0.78*	1.02**
	(0.32)	(0.32)	(0.33)	(0.33)	(0.34)	(0.39)	(0.39)	(0.46)
Flesch-Kincaid	0.16*	0.16*	0.16*	0.17*	0.19**	0.20*	0.23**	0.24**
	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)	(0.10)	(0.10)	(0.11)
Gunning Fog	0.26***	0.26***	0.26***	0.27***	0.29***	0.30**	0.30***	0.32**
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.12)	(0.11)	(0.13)
SMOG	0.17**	0.17**	0.17**	0.17**	0.19***	0.18**	0.17**	0.21**
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.09)
Dale-Chall	0.08**	0.08**	0.08**	0.08**	0.09**	0.08*	0.08*	0.10*
	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)
Editor effects	✓	✓	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓	✓	✓
Institution effects				✓	✓	✓	✓	✓
Quality controls					✓[1]	✓[1]	✓[1]	✓[1]
Native speaker					✓	✓	✓	✓
JEL (primary) effects							✓	
JEL (tertiary) effects								✓

Notes. 8,804 articles in (1)–(5); 4,917 articles in (6) and (7); 5,405 articles—including 488 from *AER Papers & Proceedings* (see Footnote 53)—in (8). Estimates are identical to those in Table 3, except that female ratio has been replaced with a dummy variable equal to 1 if a weak majority (50% or more) of authors are female. (Papers with a minority—but positive—number of female authors are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.2: Table 4, majority female-authored

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Female-authored (women)	1.87**	0.30**	0.42**	0.28**	0.12
	(0.76)	(0.15)	(0.19)	(0.14)	(0.08)
Female-authored (men)	0.60	0.12	0.11	0.05	0.03
	(0.66)	(0.13)	(0.17)	(0.12)	(0.05)
Female-authored×male	−1.26	−0.18	−0.31	−0.23	−0.09
	(1.05)	(0.20)	(0.24)	(0.18)	(0.09)
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	−20.17	−15.89	−16.95	−19.78	−21.15
Order 2	0.58	−0.32	0.11	0.29	−0.52
N_j	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓
Quality controls	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]
Native speaker	✓	✓	✓	✓	✓

Notes. Sample 9,188 observations (2,828 authors). Columns display estimates identical to those in Table 4, except that female ratio has been replaced with a dummy variable equal to 1 if a weak majority (50% or more) of authors are female. Otherwise, it is 0. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.3: Table 6, majority female-authored

	OLS	FGLS			OLS
	Published article	Working paper	Published article	Difference	Change in score
Flesch Reading Ease	1.33***	1.39**	2.48***	1.09***	1.07**
	(0.43)	(0.68)	(0.72)	(0.41)	(0.42)
Flesch-Kincaid	0.43***	0.17	0.57***	0.39***	0.39***
	(0.14)	(0.17)	(0.16)	(0.12)	(0.12)
Gunning Fog	0.45***	0.23	0.63***	0.40***	0.39***
	(0.17)	(0.18)	(0.18)	(0.13)	(0.13)
SMOG	0.28**	0.16	0.40***	0.25***	0.24***
	(0.11)	(0.11)	(0.12)	(0.09)	(0.09)
Dale-Chall	0.14***	0.17***	0.28***	0.12***	0.11***
	(0.03)	(0.05)	(0.06)	(0.03)	(0.03)
Editor effects	✓	✓	✓		✓
Journal effects	✓	✓	✓		✓
Year effects	✓	✓	✓		
Journal×Year effects	✓	✓	✓		✓
Quality controls	✓[2]	✓[2]	✓[2]		✓[3]
Native speaker	✓	✓	✓		✓

Notes. Sample 1,566 NBER working papers; 1,564 published articles (235 female-authored). Columns display estimates identical to those in Table 6, except that female ratio has been replaced with a dummy variable equal to 1 if a weak majority (50% or more) of authors are female. (Papers with a minority—but positive—number of female authors are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.4: Table 10, majority female-authored

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Majority female	3.454***	3.878***	3.874***	3.523***	3.884***	5.626***	5.658***
	(1.212)	(1.208)	(1.201)	(1.190)	(1.204)	(1.481)	(1.463)
Max. t_j	−0.219***	−0.226***	−0.222***	−0.220***	−0.220***	−0.234***	−0.225***
	(0.062)	(0.063)	(0.062)	(0.062)	(0.062)	(0.074)	(0.073)
No. pages	0.202***	0.203***	0.202***	0.202***	0.202***	0.242***	0.227***
	(0.027)	(0.027)	(0.028)	(0.027)	(0.027)	(0.041)	(0.043)
N	1.139***	1.123***	1.112***	1.133***	1.119***	1.445**	1.260**
	(0.405)	(0.402)	(0.404)	(0.405)	(0.404)	(0.554)	(0.581)
Order	0.207**	0.205**	0.203**	0.206**	0.203**	0.456**	0.435**
	(0.089)	(0.090)	(0.090)	(0.089)	(0.090)	(0.211)	(0.209)
No. citations (asinh)	−0.354*	−0.373*	−0.360*	−0.353*	−0.370*	−0.623	−0.587
	(0.197)	(0.196)	(0.197)	(0.197)	(0.196)	(0.479)	(0.474)
Mother			−4.203*		−8.119***	−15.306***	−15.546***
			(2.344)		(2.666)	(4.217)	(3.550)
Birth				−1.198	6.559	12.931*	13.580**
				(3.215)	(4.185)	(6.563)	(6.010)
Editor effects	✓	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓	✓	✓
JEL (primary) effects							✓
No. observations	2,546	2,531	2,546	2,546	2,546	1,214	1,214

Notes. Sample 2,626 articles. Columns display estimates identical to those in Table 10, except that female ratio has been replaced with a dummy variable equal to 1 if a weak majority (50% or more) of authors are female. (Papers with a minority—but positive—number of female authors are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.5: Table H.2, majority female-authored

	$t = 1$	$t = 2$	$t = 3$	$t = 4$–5	$t \geq 6$	All
Flesch Reading Ease	0.63	1.71***	3.71***	1.57	2.88**	1.94***
	(0.50)	(0.59)	(1.02)	(1.40)	(1.39)	(0.65)
Flesch-Kincaid	0.10	0.20	0.66***	0.44*	0.47	0.31**
	(0.12)	(0.16)	(0.17)	(0.24)	(0.32)	(0.14)
Gunning Fog	0.23*	0.31*	0.81***	0.55*	0.62	0.43***
	(0.13)	(0.19)	(0.21)	(0.28)	(0.38)	(0.17)
SMOG	0.15*	0.23*	0.51***	0.38*	0.45	0.29**
	(0.09)	(0.13)	(0.16)	(0.22)	(0.27)	(0.12)
Dale-Chall	0.09**	0.09	0.25**	0.16	0.29	0.16**
	(0.05)	(0.06)	(0.13)	(0.11)	(0.18)	(0.06)
No. observations	6,403	2,680	1,557	1,778	2,580	9,593
N_j	✓	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓
Journal×Year effects						✓
Institution effects	✓	✓	✓	✓	✓	✓
Quality controls	✓[4]	✓[4]	✓[4]	✓[4]	✓[4]	✓[1]
Native speaker	✓	✓	✓	✓	✓	✓

Notes. Columns display estimates identical to those in Table H.2, except that female ratio has been replaced with a dummy variable equal to 1 if a weak majority (50% or more) of authors are female. (Papers with a minority—but positive—number of female authors are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

J.2 At least one female author

TABLE J.6: Table 3, at least one female author

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Flesch Reading Ease	0.49	0.49	0.47	0.49	0.59	0.27	0.25	0.24
	(0.38)	(0.38)	(0.39)	(0.39)	(0.39)	(0.44)	(0.43)	(0.46)
Flesch-Kincaid	0.10	0.10	0.11	0.11	0.13	0.10	0.11	0.08
	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)	(0.09)	(0.09)
Gunning Fog	0.19**	0.19**	0.20**	0.21**	0.22**	0.18	0.16	0.11
	(0.09)	(0.09)	(0.10)	(0.09)	(0.10)	(0.12)	(0.10)	(0.11)
SMOG	0.12*	0.12*	0.12*	0.13*	0.14*	0.10	0.08	0.06
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)
Dale-Chall	0.07**	0.07**	0.07**	0.07**	0.08**	0.07*	0.06	0.05
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Editor effects	✓	✓	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓	✓	✓
Institution effects				✓	✓	✓	✓	✓
Quality controls					✓[1]	✓[1]	✓[1]	✓[1]
Native speaker					✓	✓	✓	✓
JEL (primary) effects							✓	
JEL (tertiary) effects								✓

Notes. 9,122 articles in (1)–(5); 5,216 articles in (6) and (7); 5,777 articles—including 561 from *AER Papers & Proceedings* (see Footnote 53)—in (8). Estimates are identical to those in Table 3, except that female ratio has been replaced with a dummy variable equal to 1 if at least one author on a paper is female. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.7: Table 4, at least one female author

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Female-authored (women)	1.31*	0.21*	0.30*	0.21*	0.12*
	(0.67)	(0.12)	(0.16)	(0.12)	(0.07)
Female-authored (men)	0.11	0.02	0.04	0.03	0.06
	(0.66)	(0.12)	(0.14)	(0.10)	(0.05)
Female-authored×male	−1.19*	−0.19	−0.26	−0.18	−0.06
	(0.64)	(0.14)	(0.16)	(0.12)	(0.07)
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	−20.14	−15.88	−16.95	−19.78	−20.98
Order 2	0.55	−0.35	0.09	0.28	−0.52
N_j	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓
Quality controls	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]
Native speaker	✓	✓	✓	✓	✓

Notes. Sample 9,188 observations (2,828 authors). Columns display estimates identical to those in Table 4, except that female ratio has been replaced with a dummy variable equal to 1 if at least one author on a paper is female. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.8: Table 6, at least one female author

	OLS	FGLS			OLS
	Published article	Working paper	Published article	Difference	Change in score
Flesch Reading Ease	0.70**	1.23**	1.72**	0.49*	0.46*
	(0.34)	(0.63)	(0.71)	(0.27)	(0.28)
Flesch-Kincaid	0.28***	0.24*	0.47***	0.22**	0.22**
	(0.10)	(0.13)	(0.15)	(0.09)	(0.09)
Gunning Fog	0.27**	0.31**	0.51***	0.20**	0.20**
	(0.11)	(0.15)	(0.17)	(0.09)	(0.09)
SMOG	0.18**	0.19*	0.33***	0.14**	0.13**
	(0.08)	(0.10)	(0.12)	(0.06)	(0.06)
Dale-Chall	0.10***	0.17***	0.24***	0.07***	0.07***
	(0.02)	(0.05)	(0.06)	(0.02)	(0.02)
Editor effects	✓	✓	✓		✓
Journal effects	✓	✓	✓		✓
Year effects	✓	✓	✓		
Journal×Year effects	✓	✓	✓		✓
Quality controls	✓[2]	✓[2]	✓[2]		✓[3]
Native speaker	✓	✓	✓		✓

Notes. Sample 1,709 NBER working papers; 1,707 published articles (378 female-authored). Columns display estimates identical to those in Table 6, except that female ratio has been replaced with a dummy variable equal to 1 if at least one author on a paper is female. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.9: Table 10, at least one female author

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1+ female	2.765***	3.022***	3.021***	2.790***	3.031***	4.150***	4.166***
	(0.964)	(0.968)	(0.962)	(0.966)	(0.963)	(1.108)	(1.084)
Max. t_j	−0.183***	−0.189***	−0.185***	−0.184***	−0.184***	−0.189**	−0.195**
	(0.060)	(0.061)	(0.060)	(0.061)	(0.061)	(0.073)	(0.073)
No. pages	0.198***	0.198***	0.197***	0.198***	0.197***	0.237***	0.223***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.040)	(0.041)
N	0.965**	0.935**	0.925**	0.961**	0.932**	1.087*	0.987
	(0.447)	(0.439)	(0.441)	(0.443)	(0.441)	(0.633)	(0.670)
Order	0.219**	0.219**	0.217**	0.219**	0.217**	0.485**	0.466**
	(0.086)	(0.087)	(0.087)	(0.087)	(0.087)	(0.191)	(0.199)
No. citations (asinh)	−0.332	−0.349*	−0.337	−0.331	−0.347*	−0.534	−0.552
	(0.203)	(0.202)	(0.202)	(0.203)	(0.201)	(0.491)	(0.492)
Mother			−3.589		−7.645***	−15.254***	−15.160***
			(2.545)		(2.508)	(2.874)	(2.655)
Birth				−0.609	6.789	14.176**	13.919**
				(3.440)	(4.070)	(5.767)	(5.606)
Editor effects	✓	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓	✓	✓
JEL (primary) effects							✓
No. observations	2,625	2,610	2,625	2,625	2,625	1,281	1,281

Notes. Sample 2,626 articles. Columns display estimates identical to those in Table 10, except that female ratio has been replaced with a dummy variable equal to 1 if at least one author on a paper is female. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.10: Table H.2, at least one female author

	$t = 1$	$t = 2$	$t = 3$	$t = 4$–5	$t \geq 6$	All
Flesch Reading Ease	0.43	0.77	3.47***	1.52	1.70	0.85
	(0.44)	(0.58)	(0.91)	(1.17)	(1.52)	(0.55)
Flesch-Kincaid	0.09	−0.03	0.61***	0.44*	0.33	0.09
	(0.10)	(0.16)	(0.16)	(0.23)	(0.34)	(0.12)
Gunning Fog	0.20*	0.08	0.72***	0.56**	0.43	0.20
	(0.11)	(0.18)	(0.19)	(0.28)	(0.41)	(0.14)
SMOG	0.14*	0.07	0.47***	0.39*	0.35	0.16
	(0.08)	(0.12)	(0.14)	(0.21)	(0.29)	(0.10)
Dale-Chall	0.09**	0.04	0.28***	0.13	0.18	0.10**
	(0.04)	(0.06)	(0.09)	(0.09)	(0.15)	(0.05)
No. observations	6,877	2,828	1,675	1,908	2,777	12,016
N_j	✓	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓
Journal×Year effects						✓
Institution effects	✓	✓	✓	✓	✓	✓
Quality controls	✓[4]	✓[4]	✓[4]	✓[4]	✓[4]	✓[1]
Native speaker	✓	✓	✓	✓	✓	✓

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

J.3 Exclusively female-authored

TABLE J.11: Table 3, 100% female-authored

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Flesch Reading Ease	0.49	0.42	0.34	0.45	0.64	0.43	0.51	0.63
	(0.57)	(0.56)	(0.57)	(0.57)	(0.60)	(0.68)	(0.69)	(0.92)
Flesch-Kincaid	0.11	0.09	0.08	0.10	0.12	0.17	0.21	0.19
	(0.13)	(0.13)	(0.13)	(0.13)	(0.15)	(0.15)	(0.15)	(0.19)
Gunning Fog	0.21	0.21	0.20	0.23	0.26	0.35**	0.36**	0.32
	(0.14)	(0.14)	(0.14)	(0.14)	(0.15)	(0.17)	(0.17)	(0.20)
SMOG	0.15	0.14	0.14	0.15	0.18	0.22*	0.22*	0.20
	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)	(0.12)	(0.12)	(0.15)
Dale-Chall	0.06	0.06	0.05	0.06	0.08	0.13**	0.12*	0.15**
	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)
Editor effects	✓	✓	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓	✓	✓
Institution effects				✓	✓	✓	✓	✓
Quality controls					✓[1]	✓[1]	✓[1]	✓[1]
Native speaker					✓	✓	✓	✓
JEL (primary) effects							✓	
JEL (tertiary) effects								✓

Notes. 8,262 articles in (1)–(5); 4,458 articles in (6) and (7); 4,841 articles—including 383 from *AER Papers & Proceedings* (see Footnote 53)—in (8). Estimates are identical to those in Table 3, except that female ratio has been replaced with a dummy variable equal to 1 if all authors on a paper are female. (Papers written by authors of both genders are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.12: Table 4, 100% female-authored

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Female-authored (women)	1.93*	0.21	0.56**	0.45**	0.31***
	(1.06)	(0.24)	(0.27)	(0.20)	(0.09)
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	−20.12	−15.88	−16.95	−19.78	−20.93
Order 2	0.56	−0.34	0.07	0.30	−0.53
N_j	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓
Quality controls	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]
Native speaker	✓	✓	✓	✓	✓

Notes. Sample 9,188 observations (2,828 authors). Estimates identical to those in Table 4, except that female ratio has been replaced with a dummy variable equal to 1 if all authors on a paper are female. Otherwise, it is 0. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Table J.13: Table 6, 100% female-authored

	OLS	FGLS			OLS
	Published article	Working paper	Published article	Difference	Change in score
Flesch Reading Ease	0.31	2.30	2.23	−0.07	−0.02
	(0.91)	(1.42)	(1.72)	(0.90)	(0.95)
Flesch-Kincaid	0.32	0.23	0.50	0.26	0.27
	(0.29)	(0.25)	(0.39)	(0.29)	(0.31)
Gunning Fog	0.33	0.40	0.64	0.24	0.25
	(0.28)	(0.30)	(0.41)	(0.30)	(0.32)
SMOG	0.15	0.39	0.45*	0.06	0.07
	(0.16)	(0.24)	(0.27)	(0.19)	(0.19)
Dale-Chall	0.11	0.38**	0.43***	0.05	0.05
	(0.08)	(0.16)	(0.15)	(0.08)	(0.09)
Editor effects	✓	✓	✓		✓
Journal effects	✓	✓	✓		✓
Year effects	✓	✓	✓		
Journal×Year effects	✓	✓	✓		✓
Quality controls	✓[2]	✓[2]	✓[2]		✓[3]
Native speaker	✓	✓	✓		✓

Notes. Sample 1,385 NBER working papers; 1,383 published articles (54 female-authored). Columns display estimates identical to those in Table 6, except that female ratio has been replaced with a dummy variable equal to 1 if all authors on a paper are female. (Papers written by authors of both genders are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.14: Table 10, 100% female-authored

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Exclusively female	6.117*	9.053**	9.054**	6.973**	9.076**	10.574**	11.143**
	(3.093)	(3.768)	(3.762)	(3.411)	(3.769)	(5.098)	(4.834)
Max. t_j	−0.236***	−0.238***	−0.234***	−0.237***	−0.233***	−0.250***	−0.241***
	(0.066)	(0.067)	(0.066)	(0.066)	(0.067)	(0.082)	(0.080)
No. pages	0.209***	0.207***	0.206***	0.208***	0.206***	0.249***	0.238***
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.038)	(0.038)
N	1.194***	1.216***	1.205***	1.193***	1.213***	1.477**	1.302**
	(0.428)	(0.426)	(0.430)	(0.429)	(0.428)	(0.578)	(0.613)
Order	0.189**	0.185**	0.183**	0.187**	0.183**	0.418*	0.388*
	(0.089)	(0.089)	(0.089)	(0.088)	(0.089)	(0.214)	(0.211)
No. citations (asinh)	−0.375*	−0.400**	−0.387*	−0.371*	−0.397**	−0.716	−0.642
	(0.194)	(0.193)	(0.193)	(0.194)	(0.193)	(0.464)	(0.467)
Mother			−9.355**		−13.327***	−20.728***	−21.850***
			(3.637)		(4.477)	(7.136)	(6.194)
Birth				−4.641	6.633	13.457*	14.498**
				(3.757)	(4.215)	(6.920)	(6.279)
Editor effects	✓	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓	✓	✓
JEL (primary) effects							✓
No. observations	2,445	2,430	2,445	2,445	2,445	1,142	1,142

Notes. Sample 2,626 articles. Columns display estimates identical to those in Table 10, except that female ratio has been replaced with a dummy variable equal to 1 if all authors on a paper are female. (Papers written by authors of both genders are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.15: Table H.2, 100% female-authored

	$t=1$	$t=2$	$t=3$	$t=4$–5	$t\geq 6$	All
Flesch Reading Ease	−0.03	2.01*	4.57***	2.59	2.50	2.96**
	(0.79)	(1.08)	(1.38)	(2.49)	(3.18)	(1.19)
Flesch-Kincaid	0.03	0.31	0.80**	0.19	0.48	0.47
	(0.18)	(0.24)	(0.32)	(0.56)	(0.46)	(0.29)
Gunning Fog	0.11	0.54**	1.23***	0.54	0.90**	0.70**
	(0.21)	(0.27)	(0.43)	(0.61)	(0.45)	(0.35)
SMOG	0.05	0.38**	0.85***	0.59	0.65**	0.49**
	(0.15)	(0.19)	(0.31)	(0.48)	(0.28)	(0.23)
Dale-Chall	0.00	0.21**	0.47***	0.50*	0.62***	0.23*
	(0.08)	(0.09)	(0.14)	(0.26)	(0.23)	(0.12)
No. observations	5,877	2,450	1,453	1,643	2,387	8,085
N_j	✓	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓
Journal×Year effects						✓
Institution effects	✓	✓	✓	✓	✓	✓
Quality controls	✓[4]	✓[4]	✓[4]	✓[4]	✓[4]	✓[1]
Native speaker	✓	✓	✓	✓	✓	✓

Notes. Columns display estimates identical to those in Table H.2, except that female ratio has been replaced with a dummy variable equal to 1 if all authors on a paper are female. (Papers written by authors of both genders are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

K Section 4.1, suplemental output

Table K.1 shows the coefficients on the journal dummies in column (2), Table 3. They compare *AER*'s readability to the readability of *Econometrica*, *JPE* and *QJE*.

Table K.1: 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***	−3.43***	−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 3. 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.

L Section 4.3, suplemental output

L.1 Table 6 (first column), full output

Table L.1 displays coefficients from estimating Equation (2) using OLS. The first row displays coefficients on working paper score (R_{jW}); the second row shows the coefficient on female ratio (β_{1P}), which is also shown in the first column of Table 6. Remaining rows present estimated coefficients on the other (non-fixed effects) control variables: max. t_j and max. T_j—*i.e.*, contemporaneous and lifetime publication counts for article j's most prolific co-author, respectively—number of citations (asinh) and a dummy variable equal to one if article j is authored by at least one native English speaker.

TABLE L.1: Table 6 (first column), full output

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
R_{jW}	0.834***	0.756***	0.773***	0.791***	0.841***
	(0.022)	(0.038)	(0.036)	(0.028)	(0.016)
Female ratio	1.347**	0.521***	0.520***	0.309**	0.179***
	(0.573)	(0.175)	(0.184)	(0.126)	(0.052)
Max. t_j	0.010	0.006	0.009	0.006	−0.004
	(0.072)	(0.018)	(0.019)	(0.012)	(0.004)
Max. T_j	0.013	0.002	0.001	0.000	0.003
	(0.054)	(0.012)	(0.013)	(0.008)	(0.003)
No. citations (asinh)	−0.320*	−0.060	−0.062	−0.051*	−0.006
	(0.181)	(0.038)	(0.044)	(0.029)	(0.015)
Native speaker	−0.222	0.011	0.027	−0.010	−0.039
	(0.426)	(0.147)	(0.192)	(0.116)	(0.027)
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Year×Journal effects	✓	✓	✓	✓	✓

Notes. Sample 1,709 NBER working papers; 1,707 published articles. Estimates exclude 279 pre-internet double-blind reviewed articles (see Footnote 65). 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 6. Coefficients on quality controls (citation counts (asinh), 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.

L.2 Table 6, accounting for field

As argued in Section 4.3 (see Footnote 60) using the change in score as the dependent variable washes out any impact field may have on an article's readability.[33] Moreover, these results—reported in the final column of Table 6—are almost identical to FGLS estimates—shown in the penultimate column—suggesting the latter are not biased by excluding them, either.

For added robustness, however, I include them here. Table L.2 replicates the FGLS (and biased OLS) estimates with dummy variables for each primary *JEL* category. As expected, figures are similar to—but standard errors higher than—those presented in Table 6.

Table L.2: Table 6, FGLS estimates controlling for *JEL* category

	OLS	FGLS		
	Published article	Working paper	Published article	Difference
Flesch Reading Ease	1.32**	2.80***	3.69***	0.88
	(0.58)	(1.05)	(1.18)	(0.59)
Flesch-Kincaid	0.55***	0.46**	0.90***	0.43**
	(0.18)	(0.23)	(0.30)	(0.20)
Gunning Fog	0.50***	0.53**	0.92***	0.39*
	(0.17)	(0.24)	(0.32)	(0.21)
SMOG	0.29**	0.38***	0.59***	0.21
	(0.11)	(0.15)	(0.19)	(0.13)
Dale-Chall	0.15***	0.33***	0.42***	0.10*
	(0.05)	(0.10)	(0.10)	(0.05)
Editor effects	✓	✓	✓	
Journal effects	✓	✓	✓	
Year effects	✓	✓	✓	
Journal×Year effects	✓	✓	✓	
Quality controls	✓[2]	✓[2]	✓[2]	
Native speaker	✓	✓	✓	
JEL (primary) effects	✓	✓	✓	

Notes. Sample 1,505 NBER working papers; 1,503 published articles. Estimates exclude 198 pre-internet double-blind reviewed articles (see Footnote 65). Columns display estimates identical to those in the first four columns of Table 6, except that fixed effects for primary *JEL* categories are included. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

[33] This assumes that field only impacts the readability of a paper when it is first drafted—*e.g.*, if concepts in certain areas are easier to explain. Thus, the *change* in readability between versions is independent of field.

M Section 4.4, supplemental output

M.1 Co-variate balance

Table M.1 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 1,554 male authors with three or more publications. The first column of the second panel ("Post-match means") displays (weighted) averages for the 110 male authors matched with a female author. Table M.2, Table M.3 and Table M.4 compare co-variate balance when restricted to matched pairs with $\underline{D}_{ik} \neq 0$.

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

TABLE M.1: Pre- and post-matching summary statistics

		Pre-match means			Post-match means		
	Women	Men	Difference	t	Men	Difference	t
T	4.55	5.89	−1.35	−3.46	4.00	0.55	2.18
$t = 1$ inst. rank	15.23	18.48	−3.25	−1.87	16.61	−1.38	−0.58
Max. citations	267.07	406.37	−139.30	−1.78	202.89	64.17	1.73
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	0.00	0.00	−0.15
1970–79	0.01	0.11	−0.09	−4.72	0.02	0.00	−0.25
1980–89	0.08	0.18	−0.10	−4.36	0.10	−0.02	−0.73
1990–99	0.19	0.21	−0.02	−0.99	0.18	0.00	0.10
2000–09	0.41	0.26	0.15	5.87	0.41	0.00	0.07
2010–15	0.31	0.20	0.11	4.20	0.29	0.02	0.40
Fraction of articles per journal							
AER	0.39	0.25	0.14	5.52	0.37	0.01	0.37
Econometrica	0.17	0.34	−0.17	−5.12	0.24	−0.07	−1.69
JPE	0.18	0.24	−0.07	−2.62	0.17	0.00	0.11
QJE	0.27	0.17	0.10	4.79	0.22	0.05	1.61
A General	0.04	0.02	0.02	1.59	0.04	0.00	0.00
B Methodology	0.00	0.02	−0.02	−1.44	0.00	0.00	
C Quant. methods	0.64	0.81	−0.17	−1.03	0.58	0.07	0.39
D Microeconomics	1.64	1.79	−0.16	−0.69	1.57	0.07	0.31
E Macroeconomics	0.58	0.62	−0.05	−0.37	0.54	0.04	0.30
F International	0.39	0.31	0.08	0.82	0.31	0.07	0.61
G Finance	0.60	0.52	0.07	0.67	0.42	0.17	1.20
H Public	0.45	0.36	0.10	1.09	0.31	0.15	1.75
I Health, welfare, edu	0.88	0.34	0.53	5.36	0.45	0.42	2.45
J Labour	1.26	0.76	0.49	3.39	0.79	0.46	2.47
K Law and econ	0.20	0.14	0.06	1.14	0.13	0.07	1.03
L Industrial org	0.73	0.57	0.16	1.47	0.53	0.20	1.52
M Marketing/acct	0.17	0.13	0.04	0.93	0.12	0.04	0.65
N Economic history	0.29	0.14	0.15	2.74	0.20	0.09	0.97
O Development	0.86	0.52	0.34	2.59	0.58	0.28	1.66
P Economic systems	0.08	0.09	−0.01	−0.22	0.05	0.03	0.78
Q Agri., environment	0.18	0.12	0.06	1.20	0.10	0.08	1.36
R Regional, transport	0.17	0.16	0.01	0.16	0.07	0.10	2.22
Z Special topics	0.16	0.10	0.06	1.50	0.15	0.01	0.14

Notes. Sample restricted to authors with three or more publications. First panel shows pre-match summary statistics (121 female authors, 1,554 male authors). Second panel shows post-match summary statistics (110 male authors). t-values for differences reported in each panel's final column.

TABLE M.2: 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	3.95	4.28	−0.34	−1.40	3.92	4.18	−0.25	−0.90
$t = 1$ inst. rank	15.36	16.87	−1.51	−0.55	16.65	15.80	0.85	0.30
Max. citations	205.39	221.80	−16.41	−0.53	223.45	245.20	−21.75	−0.48
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	−0.20	0.00	0.01	0.00	−0.21
1970–79	0.02	0.02	0.00	0.31	0.01	0.01	0.00	−0.01
1980–89	0.08	0.08	0.00	−0.17	0.09	0.08	0.00	0.13
1990–99	0.17	0.20	−0.02	−0.58	0.17	0.19	−0.03	−0.73
2000–09	0.42	0.43	−0.02	−0.38	0.40	0.40	0.00	0.02
2010–15	0.31	0.26	0.04	0.92	0.33	0.30	0.03	0.52
Fraction of articles per journal								
AER	0.38	0.38	−0.01	−0.17	0.41	0.40	0.01	0.20
Econometrica	0.24	0.19	0.05	1.08	0.23	0.19	0.04	0.91
JPE	0.15	0.19	−0.04	−1.31	0.13	0.18	−0.05	−1.44
QJE	0.24	0.24	0.00	−0.03	0.23	0.24	−0.01	−0.20
A General	0.07	0.07	−0.01	−0.14	0.03	0.02	0.01	0.35
B Methodology	0.00	0.00	0.00		0.00	0.00	0.00	
C Quant. methods	0.54	0.47	0.07	0.39	0.51	0.49	0.02	0.12
D Microeconomics	1.63	1.60	0.03	0.13	1.67	1.66	0.02	0.06
E Macroeconomics	0.48	0.47	0.01	0.05	0.43	0.47	−0.04	−0.25
F International	0.27	0.34	−0.07	−0.48	0.26	0.29	−0.03	−0.25
G Finance	0.42	0.51	−0.08	−0.45	0.43	0.59	−0.15	−0.84
H Public	0.34	0.40	−0.06	−0.67	0.41	0.40	0.01	0.12
I Health, welfare, edu	0.53	0.81	−0.28	−1.33	0.65	0.79	−0.14	−0.62
J Labour	0.86	1.11	−0.25	−1.17	0.91	1.07	−0.16	−0.74
K Law and econ	0.12	0.16	−0.05	−0.66	0.14	0.21	−0.07	−0.87
L Industrial org	0.62	0.66	−0.04	−0.24	0.57	0.76	−0.20	−1.25
M Marketing/acct	0.12	0.13	−0.01	−0.14	0.12	0.13	−0.01	−0.15
N Economic history	0.23	0.38	−0.15	−1.17	0.20	0.31	−0.11	−0.93
O Development	0.53	0.59	−0.06	−0.37	0.49	0.59	−0.10	−0.68
P Economic systems	0.05	0.09	−0.04	−0.75	0.05	0.11	−0.05	−0.92
Q Agri., environment	0.10	0.14	−0.04	−0.71	0.04	0.09	−0.05	−1.19
R Regional, transport	0.09	0.12	−0.03	−0.61	0.10	0.15	−0.06	−1.02
Z Special topics	0.17	0.18	0.00	−0.03	0.15	0.18	−0.02	−0.35

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

TABLE M.3: 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	3.84	4.06	−0.22	−0.92	3.84	4.09	−0.25	−1.07
$t = 1$ inst. rank	16.91	18.39	−1.48	−0.50	15.20	17.90	−2.70	−0.90
Max. citations	201.51	243.87	−42.35	−1.21	205.39	230.70	−25.32	−0.70
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	−0.22	0.00	0.01	0.00	−0.18
1970–79	0.02	0.02	0.00	0.16	0.02	0.02	0.00	0.07
1980–89	0.11	0.10	0.01	0.17	0.09	0.08	0.01	0.39
1990–99	0.16	0.17	−0.01	−0.31	0.16	0.17	−0.01	−0.35
2000–09	0.37	0.39	−0.02	−0.45	0.39	0.41	−0.02	−0.41
2010–15	0.34	0.31	0.03	0.52	0.34	0.32	0.02	0.38
Fraction of articles per journal								
AER	0.40	0.38	0.02	0.41	0.39	0.39	−0.01	−0.12
Econometrica	0.21	0.20	0.00	0.08	0.21	0.19	0.02	0.38
JPE	0.16	0.19	−0.03	−0.88	0.15	0.19	−0.04	−1.06
QJE	0.23	0.22	0.01	0.19	0.25	0.22	0.02	0.61
A General	0.02	0.02	0.00	−0.09	0.05	0.04	0.01	0.31
B Methodology	0.00	0.00	0.00		0.00	0.00	0.00	
C Quant. methods	0.42	0.48	−0.05	−0.38	0.47	0.47	0.00	0.01
D Microeconomics	1.48	1.44	0.04	0.17	1.47	1.56	−0.08	−0.37
E Macroeconomics	0.52	0.48	0.05	0.30	0.52	0.49	0.02	0.15
F International	0.29	0.24	0.04	0.38	0.26	0.22	0.04	0.31
G Finance	0.39	0.52	−0.14	−0.77	0.42	0.48	−0.06	−0.32
H Public	0.43	0.41	0.02	0.17	0.42	0.41	0.02	0.14
I Health, welfare, edu	0.58	0.83	−0.25	−1.14	0.56	0.88	−0.31	−1.40
J Labour	0.94	1.11	−0.17	−0.76	0.96	1.12	−0.16	−0.71
K Law and econ	0.11	0.17	−0.06	−0.93	0.13	0.17	−0.04	−0.64
L Industrial org	0.44	0.59	−0.14	−0.97	0.56	0.54	0.02	0.14
M Marketing/acct	0.09	0.10	−0.01	−0.18	0.09	0.10	0.00	−0.09
N Economic history	0.29	0.41	−0.13	−0.95	0.29	0.43	−0.14	−1.00
O Development	0.48	0.57	−0.10	−0.69	0.55	0.49	0.06	0.43
P Economic systems	0.06	0.10	−0.04	−0.74	0.04	0.09	−0.05	−1.04
Q Agri., environment	0.07	0.11	−0.04	−0.81	0.07	0.11	−0.04	−0.73
R Regional, transport	0.06	0.15	−0.09	−1.94	0.08	0.14	−0.05	−1.06
Z Special topics	0.14	0.15	0.00	−0.03	0.15	0.16	−0.01	−0.10

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

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

	Dale-Chall			
	Discrimination			
	Against women	Against men	Difference	t
T	4.16	4.52	−0.36	−1.18
$t = 1$ inst. rank	14.13	17.08	−2.95	−1.16
Max. citations	215.18	275.44	−60.26	−1.30
Fraction of articles per decade				
1950–59	0.00	0.00	0.00	
1960–69	0.00	0.01	0.00	−0.20
1970–79	0.02	0.02	0.00	0.30
1980–89	0.09	0.08	0.01	0.23
1990–99	0.17	0.19	−0.02	−0.67
2000–09	0.41	0.43	−0.02	−0.53
2010–15	0.31	0.27	0.04	0.84
Fraction of articles per journal				
AER	0.37	0.38	−0.01	−0.27
Econometrica	0.21	0.17	0.04	0.92
JPE	0.18	0.20	−0.02	−0.53
QJE	0.23	0.24	−0.01	−0.32
A General	0.05	0.06	0.00	−0.12
B Methodology	0.00	0.00	0.00	
C Quant. methods	0.41	0.41	0.00	0.01
D Microeconomics	1.80	1.69	0.11	0.45
E Macroeconomics	0.65	0.56	0.10	0.54
F International	0.28	0.30	−0.01	−0.08
G Finance	0.47	0.59	−0.12	−0.63
H Public	0.43	0.39	0.05	0.42
I Health, welfare, edu	0.45	0.84	−0.39	−1.92
J Labour	0.92	1.31	−0.39	−1.66
K Law and econ	0.18	0.23	−0.05	−0.58
L Industrial org	0.68	0.61	0.07	0.46
M Marketing/acct	0.13	0.15	−0.02	−0.31
N Economic history	0.28	0.38	−0.09	−0.72
O Development	0.65	0.77	−0.12	−0.62
P Economic systems	0.06	0.11	−0.05	−0.89
Q Agri., environment	0.09	0.15	−0.05	−0.78
R Regional, transport	0.13	0.14	−0.01	−0.18
Z Special topics	0.16	0.19	−0.04	−0.51

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

M.2 List of authors in each matched pair

Table M.5: Matched pairs

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

Table M.5 (continued)

Matched pairs		Matched pairs	
Female	Male	Female	Male
Hastings, Justine S.	Pope, Devin G.	Williams, Heidi L.	Gowrisankaran, Gautam
Ho, Katherine	Kremer, Ilan	Wooders, Myrna Holtz	Gallant, A. Ronald
Hoxby, Caroline Minter	Kessler, Daniel P.	Yariv, Leeat	Lange, Andreas
İmrohoroğlu, Ayşe	Casari, Marco	Yellen, Janet L.	Freeman, Richard B.
Jayachandran, Seema	Caselli, Francesco	Zeiler, Kathryn	van Soest, Arthur
Kowalski, Amanda E.	Mahoney, Neale	Zhuravskaya, Ekaterina	Kuhn, Peter
Kranton, Rachel E.	Kosfeld, Michael		

Notes. Table lists the names of the matched pairs from Section 4.4.2. In each panel, female members are listed first; male members second. See Section 4.4.2 for details on the matching process.

M.3 $\widehat{R}_{it}$ regression output

Table M.6 displays output from time- and gender-specific regressions used to generate $\widehat{R}_{it}$ (Equation (14)).

TABLE M.6: Regression output generating $\widehat{R}_{it}$ (Equation (14))

	Women		Men	
	$t=1$	$t=3$	$t=1$	$t=3$
Flesch Reading Ease				
Female ratio	1.36	2.99	−4.24	8.21
	(4.16)	(3.88)	(7.96)	(5.51)
Constant	38.24***	41.17***	37.99***	37.96***
	(3.15)	(2.47)	(1.13)	(1.20)
Flesch Kincaid				
Female ratio	−0.13	0.48	0.43	2.43**
	(0.86)	(0.78)	(1.78)	(1.21)
Constant	−13.72***	−13.33***	−13.93***	−14.20***
	(0.65)	(0.50)	(0.25)	(0.26)
Gunning Fog				
Female ratio	−0.30	1.01	−0.59	2.25
	(1.04)	(0.97)	(2.04)	(1.46)
Constant	−17.15***	−17.22***	−17.70***	−17.90***
	(0.79)	(0.62)	(0.29)	(0.32)
SMOG				
Female ratio	−0.15	0.74	−0.27	1.39
	(0.76)	(0.72)	(1.41)	(1.03)
Constant	−15.07***	−15.19***	−15.54***	−15.61***
	(0.57)	(0.46)	(0.20)	(0.23)
Dale-Chall				
Female ratio	−0.06	0.48	−1.63**	0.61
	(0.35)	(0.39)	(0.78)	(0.41)
Constant	−10.96***	−11.11***	−11.08***	−11.25***
	(0.26)	(0.25)	(0.11)	(0.09)

Notes. Sample 121 female authors; 110 male authors. Sample restricted to matched authors. See Section 4.4.2 for details on how matches were made. Regressions weighted by the frequency observations are used in a match. Standard errors in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

M.4 Table 8, Equation (13) and Condition 3

Table M.7 estimates D_{ik} with Equation (13). Table M.8 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 4.4.2.

TABLE M.7: D_{ik}, Equation (13)

	Discrimination against women ($\underline{D}_{ik} > 0$)			Discrimination against men ($\underline{D}_{ik} < 0$)			Mean, all observations	
	Mean	S.D.	N	Mean	S.D.	N	(1)	(2)
Flesch Reading Ease	9.23	7.37	59	−5.43	5.12	23	3.40***	2.22**
							(0.77)	(0.89)
Flesch Kincaid	1.74	1.28	60	−1.27	1.35	23	0.64***	0.47***
							(0.15)	(0.17)
Gunning Fog	2.29	1.82	58	−1.55	1.53	23	0.84***	0.58**
							(0.20)	(0.23)
SMOG	1.86	1.38	51	−1.00	0.95	26	0.59***	0.40**
							(0.15)	(0.17)
Dale-Chall	0.90	0.64	63	−0.69	0.44	21	0.37***	0.28***
							(0.08)	(0.09)

Notes. Table displays estimates identical to those in Table 8, except that $\underline{D}_{ik}$ is determined by Equation (13). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE M.8: D_{ik}, proxying for acceptance rates (Condition 3)

	Discrimination against women ($\underline{D}_{ik} > 0$)			Discrimination against men ($\underline{D}_{ik} < 0$)			Mean, all observations	
	Mean	S.D.	N	Mean	S.D.	N	(1)	(2)
Flesch Reading Ease	13.72	10.87	37	−7.56	6.74	18	3.52***	1.96
							(1.15)	(1.29)
Flesch Kincaid	2.82	2.36	36	−2.26	2.14	21	0.51*	0.29
							(0.27)	(0.28)
Gunning Fog	3.29	3.01	39	−2.48	2.68	19	0.81***	0.49
							(0.31)	(0.34)
SMOG	2.86	2.11	32	−1.62	1.89	20	0.56**	0.32
							(0.23)	(0.25)
Dale-Chall	1.32	0.93	37	−1.13	0.73	16	0.35***	0.22
							(0.13)	(0.14)

Notes. Table displays estimates identical to those in Table 8, except that a matched pair is categorised as discrimination against i (k) only if $T_i \leq T_k$ ($T_k \leq T_i$) holds as well. Otherwise, Theorem 1 is inconclusive. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

M.5 $\widehat{R}_{it}$, *controlling for* JEL *category*

Table M.9 and Figure M.1 replicate the analysis in Section 4.4.2 but Equation (14) controls for primary *JEL* category. $\widehat{R}_{it}$ was reconstructed at female ratio equal to 1 for women, 0 for men and for a paper classified in *JEL* categories D (microeconomics) and J (labour and demographic economics).

TABLE M.9: $\underline{D}_{ik}$, controlling for *JEL* category

	Discrimination against women ($\underline{D}_{ik} > 0$)			Discrimination against men ($\underline{D}_{ik} < 0$)			Mean, all observations	
	Mean	S.D.	N	Mean	S.D.	N	(1)	(2)
Flesch Reading Ease	15.54	10.59	44	−9.48	8.79	20	5.59**	4.88*
							(2.58)	(2.68)
Flesch Kincaid	3.43	2.41	41	−2.15	1.61	11	1.39***	1.08*
							(0.51)	(0.58)
Gunning Fog	3.69	2.70	38	−2.78	2.75	19	1.11*	0.76
							(0.63)	(0.68)
SMOG	2.71	1.90	37	−2.06	1.95	20	0.73	0.50
							(0.45)	(0.49)
Dale-Chall	1.65	1.00	28	−0.99	0.62	11	0.45**	0.29
							(0.21)	(0.24)

Notes. Sample 87 matched pairs (79 and 87 distinct men and women, respectively). Table displays estimates identical to those in Table 8, except that Equation (14) includes primary *JEL* classification dummies; $\widehat{R}_{it}$ was reconstructed at female ratio equal to 1 for women, 0 for men and a paper classified in *JEL* categories D and J. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

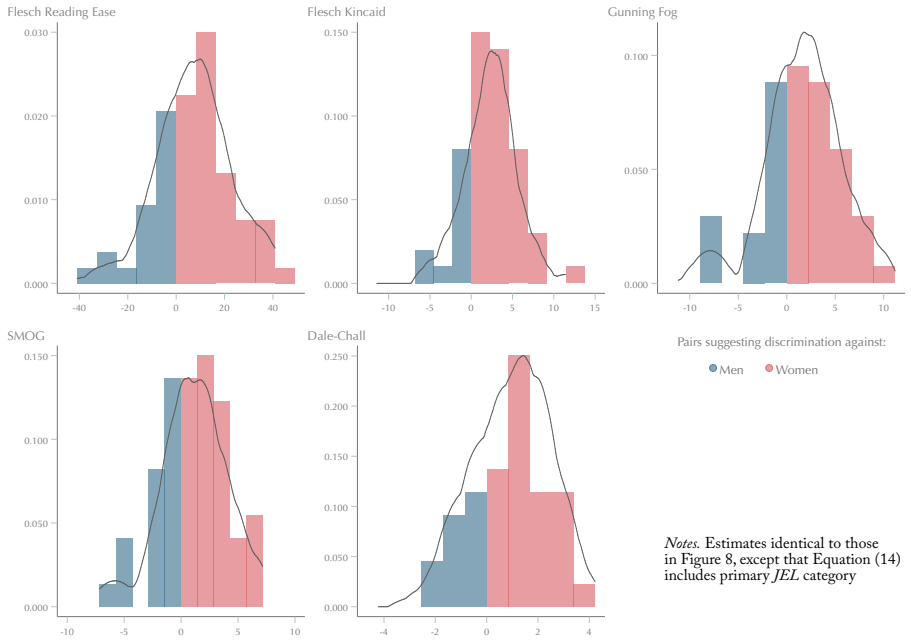

FIGURE M.1: Distributions of $\underline{D}_{ik}$, controlling for *JEL* category

N Section 4.6, supplemental output

N.1 Table 10, alternative year fixed effects

Table N.1 and Table N.2 replicate Table 10, replacing publication year fixed effects with fixed effects for submission and acceptance years, respectively. As discussed in Footnote 118, submission year effects narrow the gender gap in publication times by roughly two months; the gap marginally widens when acceptance years are used instead.

Table N.1: Table 10, submission year effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female ratio	2.667**	4.047***	4.075***	3.103**	4.090***	4.654**	4.537**
	(1.279)	(1.398)	(1.398)	(1.412)	(1.343)	(1.709)	(1.767)
Max. t_j	−0.117***	−0.121***	−0.118***	−0.118***	−0.117***	−0.123**	−0.124**
	(0.041)	(0.042)	(0.042)	(0.041)	(0.042)	(0.048)	(0.049)
No. pages	0.157***	0.155***	0.155***	0.156***	0.155***	0.160***	0.153***
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.024)	(0.027)
N	0.990***	0.938***	0.933***	0.967***	0.940***	0.854***	0.771**
	(0.240)	(0.235)	(0.238)	(0.236)	(0.273)	(0.280)	(0.300)
Order	0.137**	0.131**	0.131**	0.135**	0.131**	0.127	0.147
	(0.061)	(0.061)	(0.061)	(0.061)	(0.065)	(0.116)	(0.125)
No. citations (asinh)	−0.474***	−0.487***	−0.483***	−0.471***	−0.492***	−1.260***	−1.282***
	(0.164)	(0.165)	(0.163)	(0.163)	(0.165)	(0.275)	(0.263)
Mother			−7.284***		−10.956***	−20.729***	−20.705***
			(2.209)		(3.216)	(2.215)	(2.740)
Birth				−3.855	6.138*	17.112***	17.079***
				(3.125)	(3.506)	(2.975)	(2.991)
Editor effects	✓	✓	✓	✓	✓	✓	✓
Sub. year effects	✓	✓	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓	✓	✓
JEL (primary) effects							✓
No. observations	2,623	2,608	2,623	2,623	2,623	1,281	1,281

Notes. Sample 2,626 articles. Estimates are identical to those in Table 10 except that submission year effects are used instead of publication year effects. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Table N.2: Table 10, acceptance year effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female ratio	5.342**	6.901***	6.869***	5.773**	6.896***	9.560***	9.557***
	(2.236)	(2.372)	(2.370)	(2.316)	(2.109)	(3.149)	(3.040)
Max. t_j	−0.170***	−0.173***	−0.171***	−0.171***	−0.169***	−0.169**	−0.176**
	(0.050)	(0.051)	(0.051)	(0.051)	(0.061)	(0.064)	(0.071)
No. pages	0.198***	0.197***	0.196***	0.197***	0.196***	0.238***	0.225***
	(0.022)	(0.023)	(0.023)	(0.023)	(0.028)	(0.033)	(0.035)
N	1.125***	1.078***	1.065***	1.103***	1.074**	1.366***	1.236**
	(0.369)	(0.362)	(0.363)	(0.362)	(0.407)	(0.488)	(0.523)
Order	0.214**	0.211**	0.209**	0.213**	0.209**	0.479***	0.477***
	(0.082)	(0.082)	(0.082)	(0.082)	(0.087)	(0.163)	(0.163)
No. citations (asinh)	−0.398*	−0.422*	−0.410*	−0.396*	−0.422**	−0.663	−0.657
	(0.215)	(0.213)	(0.214)	(0.216)	(0.203)	(0.456)	(0.469)
Mother			−7.963**		−12.505***	−22.902***	−23.023***
			(3.678)		(3.335)	(3.146)	(3.332)
Birth				−3.850	7.544*	16.632***	16.357***
				(4.189)	(4.226)	(5.153)	(4.612)
Editor effects	✓	✓	✓	✓	✓	✓	✓
Accepted year effects	✓	✓	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓	✓	✓
JEL (primary) effects							✓
No. observations	2,625	2,610	2,625	2,625	2,625	1,281	1,281

Notes. Sample 2,626 articles. Estimates are identical to those in Table 10 except that acceptance year effects are used instead of publication year effects. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

N.2 Table 10, alternative thresholds for mother$_j$

Table N.3 repeats the regression presented in Table 10 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 N.3: Table 10, alternative thresholds for mother$_j$

	Age < 3	Age < 4	Age < 5	Age < 10	Age < 18
Female ratio	5.945***	6.157***	6.822***	6.695***	6.418***
	(2.112)	(2.069)	(2.092)	(2.114)	(2.157)
Mother	−4.676*	−7.966**	−11.353***	−8.655**	−4.652
	(2.361)	(3.044)	(3.083)	(3.532)	(3.577)
Birth	1.072	4.131	6.895*	4.314	0.581
	(3.789)	(3.513)	(4.084)	(4.436)	(4.133)
Max. t_j	−0.176***	−0.176***	−0.175***	−0.175***	−0.175***
	(0.062)	(0.062)	(0.062)	(0.062)	(0.062)
No. pages	0.196***	0.195***	0.195***	0.195***	0.196***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
N	1.092**	1.088**	1.067**	1.066**	1.073**
	(0.423)	(0.422)	(0.422)	(0.423)	(0.424)
Order	0.217**	0.216**	0.214**	0.214**	0.215**
	(0.086)	(0.086)	(0.087)	(0.086)	(0.086)
No. citations (asinh)	−0.336	−0.340	−0.358*	−0.352*	−0.344*
	(0.204)	(0.203)	(0.202)	(0.202)	(0.202)
Editor effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓

Notes. Sample 2,625 articles. Coefficients from OLS estimation of Equation (16) at different age thresholds for mother$_j$. In column one, mother$_j$ equals one for papers authored exclusively by women with children younger than three; in column two, the age threshold is four; *etc.* Column three corresponds to results presented in Table 10. Standard errors clustered by year in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

References

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

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.

Chall, J. S. and E. Dale (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, Massachusetts: Brookline Books.

Hengel, E. (2015). "Two Essays on Bankruptcy and One Essay on Gender Differences in Academic Publishing". PhD thesis. University of Cambridge.

— (2016). "Gender Differences in Peer Review Scrutiny". Mimeo.

Liang, F. M. (1983). "Word Hy-phen-a-tion by Com-put-er". PhD thesis. Stanford University.