Publishing while Female
Gender Differences in Peer Review Scrutiny

Erin Hengel[*]

October 2016

Using five well-known "readability" tests, I analyse every article abstract published in the top four economics journals since 1950. I) Abstracts written by women are 1–6 percent more readable than those by men. II) The gap is up to three times higher in published articles than in earlier, draft versions of the same papers. III) Women's writing gradually improves but men's does not—meaning the readability gap grows over authors' careers. I explore many interpretations; the simplest and most persuasive is that referees apply higher standards to women's writing, subjecting them to an added time tax. This last hypothesis is confirmed by submit-accept times at *Econometrica*: female-authored papers take six months longer to complete peer review.

1 Introduction

Ladies, we aren't that common in economics. Only a third, fifth and tenth of assistant, associate and full professors, respectively, are women (Romero, 2013). Female economists are less likely to make tenure, take longer when they do and earn much less than their male peers (Bandiera, 2016; Ceci et al., 2014; Ginther and Kahn, 2004; Sarsons, 2015; Weisshaar, 2014).

These statistics are uncomfortable, but their causes are myriad: lower publishing rates, career choices, motherhood and, probably, bias. In lab experiments women are subject to tougher standards. Their qualifications and ability are underestimated (Foschi, 1996; Grunspan et al., 2016; Moss-Racusin et al., 2012; Reuben et al., 2014). Female-authored manuscripts are evaluated more critically (P. Goldberg, 1968; Krawczyk and Smyk, 2016; Paludi and Bauer, 1983); when collaborating with with men, women are given less credit (Heilman and Haynes, 2005; Sarsons, 2015).

Peer review is not immune. Using five reliable measures of writing clarity, I show that female-authored articles published in top economics journals are better written than similar papers by men; the simplest interpretation is that editors and referees expect clearer, more direct writing from women. Because better writing takes effort to compose, higher standards prolong female review times—by six months at *Econometrica*—and may be a fundamental factor behind lower publishing rates.

Prior investigations suggest journal acceptance rates are genuinely bias-free (see, *e.g.*, Abrevaya and Hamermesh, 2012; Blank, 1991; Borsuk et al., 2009; Gilbert et al., 1994; Lloyd,

[*] University of Liverpool Management School. erin.hengel@gmail.com.

This paper is a revised version of the third chapter of my Ph.D. dissertation (University of Cambridge, Faculty of Economics, September 2015). I am grateful to my supervisor Christopher Harris for (a) excellent guidance and (b) thinking this was a good idea. I also owe special thanks to Gary Cook, Jeremy Edwards, Leonardo Felli, Carola Frege, Jane Hunt, Hamish Low, Brendan McCabe and Kevin Schnepel for helpful comments and to Michael Hengel (my dad) for research assistance. All errors are, of course, my own.

1990).[1] To the best of my knowledge, however, gender neutrality is established in only a narrow context (publication outcomes) using this single indicator. I ask a different question. Men's and women's papers may be published at comparable *rates*, but are they reviewed with comparable *scrutiny*? For, if women are stereotypically assumed less capable at math, logic and reasoning than men and generally need more evidence to rate as equally competent, some well-intentioned referees might (unknowingly) inspect their papers more closely, demand a larger number of revisions and, in general, be less tolerant of complicated, dense writing.

Complicated, dense writing is my focus. In the English language, more clearly written prose is better prose, all things equal. Thoughtful word choice and simple sentence structure make text easier to understand, more interesting to read and expose inconsistencies long-winded writing often hides. Journal editors tend to agree—*Econometrica* asks authors to write "crisply but clearly" and to take "the extra effort involved in revising and reworking the manuscript until it will be clear to most if not all of our readers" (*Econometrica* submission guidelines, June 2016).[2]

If referees hold female- and male-authored papers to identical standards, both should be equally well written. To test this, I rely on a relationship familiar to linguists and educators: simple vocabulary and short sentences are easier to understand and straightforward to quantify. Using the five most widely used, studied and reliable formulas to exploit this, I analyse 9,123 article abstracts[3] published in the *American Economic Review* (*AER*), *Econometrica* (*ECA*), *Journal of Political Economy* (*JPE*) and *Quarterly Journal of Economics* (*QJE*).[4]

I find systematic evidence that journal referees are more critical of women's papers. First, female-authored abstracts are 1–6 percent more readable than those by men. Women write better despite controls for editor, journal, year and *JEL* classification; that remains unchanged when proxying for article and author quality. 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; nor (iii) generated by factors correlated with gender but really related to knowledge, intelligence and creativity.

Second, the gender readability gap widens during peer review. I compare National Bureau of Economic Research (NBER) working papers to their final, published versions; the gap is up to three times larger for the latter.[5] While both papers are exposed to many factors that impact readability, only published articles are subject to peer review. By comparing the two, influences unrelated to immediate peer review are isolated from those that are; assuming the former are not correlated with the latter's timing, a widening gap suggests a causal link.[6]

Third, women's writing gradually gets better but men's does not. Between authors' first and

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

[2] 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). In a random selection of 100 posts on Shit My Reviewers Say, a quarter deal with writing quality, document structure or word choice/tone.

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

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

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

[6] The suggested causal link is with peer review, but not necessarily with referee scrutiny. This issue and the timing-independence assumption are addressed in Section 3.2.1 and Section 4.1.

third published articles, the readability gap increases by 12 percent. Evidence does not suggest senior female economists co-author with more women. Nor are initially bad female writers leaving academia. Instead, women apparently figure out that better writing makes peer review smoother; they write subsequent papers clearer from the start.

Points one to three provide strong evidence that peer review is at least partially responsible for better writing in female-authored papers. But is it given that female-authored papers invite undue scrutiny? No. In Section 4.1 I explore several alternative hypotheses—many of which have nothing to do with bias and some that exonerate peer review, too.[7] Yet as Section 4.1 also illustrates, the most straightforward—and, frankly, persuasive—explanation is indeed that journal referees are more critical of female-authored papers.

Clearer sentences, less jargon and more scrutiny aren't bad things. Papers that are easier to understand enjoy wider, more diverse readership; closer review catches logical mistakes and leads to fewer factual errors. Still, extra attention isn't costless: adding robustness checks, clarifying proofs and making sentences even marginally more readable takes time.[8]

Apparently a lot of time. As shown in Section 3.2.2, female-authored papers need ***six months longer*** to complete peer review. This estimate is based on data from *Econometrica*, controls for relevant factors—including whether the author was a mother to young children and/ or gave birth during her review—and is highly significant. Spending six more months in peer review is a serious time tax—undoubtedly hurting women's productivity and probably, as a consequence, promotion rates, too.

This paper contributes to the literature in four ways. First, to the best of my knowledge, mine is the only study to suggest and document evidence of gender bias in the peer review process (as opposed to its outcome).[9] Its findings confirm similar conclusions in research on employee performance reviews, teaching evaluations and online comments—women receive more abusive feedback, less credit for intelligence and creativity and are expected to be more organised, prepared and clear (Boring, 2015; Correll and Simard, 2016; Gardiner et al., 2016).

Second, this paper proposes a novel explanation for academia's "Publishing Paradox", "Leaky Pipeline" and general promotion gap.[10] Traditional hypotheses centre on motherhood, family commitments and gender differences in behaviour (Ceci et al., 2014; Niederle and Vesterlund,

[7] Among them: Are female economists disproportionately native English speakers? Do men ignore editorial changes suggested by referees? Are female referees the toughest critics, and if so, are they more likely to review female-authored papers? Do manuscripts written by women deserve more criticism because they aren't as good?

[8] While 1–6 percent seems small, it is based on a single paragraph. Assuming a similar standard applies to every paragraph in a paper and improving each one takes slightly more time, the accumulated impact may be substantial. See Berk et al. (2015) for a general discussion on good referee reports—and how current culture may encourage extraneous (and time-consuming) demands in otherwise publishable papers.

[9] A biased peer review process was suggested in a recent online debate following the submission experience of two female biologists at a PLOS ONE journal. The women were instructed by their referee to "find one or two male biologists to work with (or at least obtain internal peer review from, but better yet as active co-authors)" to prevent the paper from "drifting too far away from empirical evidence into ideologically biased assumptions". PLOS ONE responded, agreed to re-review the paper and proposed open peer review to deal with the general issue (Bernstein, 2015).

[10] "Publishing Paradox" and "Leaky Pipeline" refer to phenomena in academia whereby women publish fewer papers and disproportionately leave the profession, respectively. Evidence on whether female academics are hired and promoted at lower rates is mixed. One study suggests so-called STEM (science, technology, engineering, mathematics) fields actually prefer hiring women—although male economists continue to show a slight (but not significant) preference for men (Williams et al., 2015). Other studies find male candidates are preferred in postdoctoral research and laboratory management positions (Moss-Racusin et al., 2012; Sheltzer and Smith, 2014). Men are also more likely granted tenure when compared to women with an identical publication history (Weisshaar, 2014). 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).

2010). I suggest, however, that unequal time spent making revisions leads to unequal time conducting new research—and may be an important reason for lower female productivity.[11] It might also rationalise why so few female economists publish solo-authored work although it's the only work women receive full and fair credit for (Sarsons, 2015).[12]

Third, my findings emphasise the importance of transparency and monitoring. Unlike referee reports, journal acceptance rates are easy to measure and frequently audited; both factors foster accountability, which encourages gender neutrality (Foschi, 1996). Monitoring referee reports is difficult, but it isn't impossible—especially if peer review were open. As discussed in Section 4.2, several science and medical journals not only reveal referees' identities, they also post reports online. Quality does not decline (it may actually increase), referees still referee (even those who initially refuse) and the extra 25–50 minutes spent reviewing seems tolerable (van Rooyen et al., 1999; van Rooyen et al., 2010; Walsh et al., 2000).

This paper's final contribution is technical. I use readability scores to untap a largely ignored, naturally occurring source of pseudo-experiments relevant to research on gender or racial bias—and differential group treatment, more generally.[13] Readability scores have their limitations (see Section 2.1) and their use in this manner applies to just a narrow set of questions. Nevertheless, they are cheaper than audit studies and arguably more objective than survey data. An analogous approach may (or may not) expose similar group differences in, *inter alia*, successful business proposals funded by venture capitalists, letters to the editor published in newspapers or annual report introductions by CEOs.

The remainder of this paper is organised as follows. Section 2 describes the data and readability measures used in the analysis. Section 3 presents results; a detailed discussion (including possible explanations) comes next (Section 4). Section 5 concludes.

2 Data

The data include every English article published in *AER*, *Econometrica*, *JPE* and *QJE* between January 1950 and December 2015 (inclusive). Prior research has found authors write in a stylistically consistent manner across the abstract, introduction and discussion section of a peer reviewed article (Hartley et al., 2003b).[14] Of these three, I concentrate on abstracts. Abstract structure is standardised in a manner optimal for computing readability scores: 100–200 words, no citations and few abbreviations and equations (Dale and Chall, 1948). Abstracts are self-contained, universally summarise the research and are the first and most frequently read part of an article (King et al., 2006)—all factors suggesting a relatively homogenous degree of review across journals and subject matter. Conveniently, most have also been converted to accurate machine-readable text by digital libraries and bibliographic databases.

[11] See Footnote 77 for a discussion on gender differences in scientific publishing rates.

[12] Sarsons (2015) documents a female-specific co-authorship penalty in economics faculty tenure decisions—but no evidence that women co-author less to overcome it. She hypothesises that women are simply unaware of the penalty (survey data supports her idea). Alternatively, women may reasonably still co-author with men because they are willing to forgo credit in order to speed up the review process.

[13] Using readability scores to uncover gender bias in the way news is reported was first proposed by Ali et al. (2010). In an effort to determine gender differences in writing styles, Hartley et al. (2003b) compare male and female Flesch Reading Ease scores for 80 papers published in the *Journal of Educational Psychology*; they found no consistent, sex-specific difference. See Footnote 81 and Footnote 82 for a discussion and list of other creative ways readability scores have been used in academic research.

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

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	634	1	1,295
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	2,447	1,475	9,123

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.

The largest sample comes from *Econometrica* which consistently published abstracts with its articles prior to 1950. *JPE* added them in the 1960s and *QJE* in 1980. *AER* came last in 1986.[15] Table 1 displays data coverage by journal and decade. Bibliographic information and PDFs were scraped from the websites of Oxford Journals, the American Economic Association, the Econometric Society, Wiley, JSTOR and EBSCO.

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

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

In total, 1,039 different institutions were identified. I created 64 dummy variables, each of which represents one or more institution(s); groupings reflect counts of distinct articles in which an institution was listed as an affiliation.[16] Specifically, institutions listed in 59 or fewer articles were grouped in bins of 10 to form six dummy variables: the 751 institutions mentioned in 0–9 articles were grouped to form the first dummy variable, the 92 mentioned in 10–19 articles were grouped to form the second, *etc.* Fifty-eight institutions were affiliated with 60 or more articles; each is assigned its own dummy variable. When multiple institutions are associated with an observation, only the dummy variable with the highest-rank is used, *i.e.*, the highest-ranked institution per author when data is analysed at the author-level and the

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

[16] 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.

highest-ranked institution for all authors when data is analysed at the article-level.

To control for author quality/productivity, I generated 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).[17] For co-authored articles, only the dummy variable corresponding to the most prolific author is used.

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

To control for editorial policy, I recorded editor/editorial board member names from issue mastheads. *AER* and *Econometrica* employ an individual to oversee policy. *JPE* and *QJE* do not generally name one lead editor and instead rely on boards composed of four to five faculty members at the University of Chicago and Harvard, respectively.[18] Editor controls are based on distinct lead editor/editorial boards—*i.e.*, they differ by at least one member. In total, 74 groups are formed in this manner.

The analysis in Section 3.2 uses two additional datasets. The first matches published articles with NBER working papers. Matches are first attempted using citation data from RePEc and then by searching NBER's database directly for unmatched papers authored by an NBER family member. 1,978 published articles were eventually matched to 2,008 NBER working papers—approximately one-fifth of the data.[19] Bibliographic information and abstract text were scraped from www.nber.org.

The second dataset analysed in Section 3.2 compiles submit-accept times at *Econometrica*—the only journal that makes any kind of disaggregated data on the revision process publicly available.[20] I extracted this information from digitised articles using the open source command utility `pdftotext`.

To control for motherhood's impact on revision times, I recorded children's birth years for women with at least one 100 percent female-authored paper in *Econometrica*. I personally (and, I apologise, rather unsettlingly) gleaned this information from published profiles, CVs, acknowledgements, Wikipedia, personal websites, Facebook pages, intelius.com background checks and local school district/popular extra-curricular activity websites.[21] Exact years were recorded whenever found; otherwise, they were approximated by subtracting a child's actual

[17] This quality/productivity control has several limitations: (i) it relies on publication counts—not necessarily an accurate measure of "quality"; (ii) it discounts younger economists' productivity; and (iii) it generates somewhat inconsistent groupings—for example, two authors have published 45 articles, but only one author has published 37 articles (Andrei Shleifer). Appendix D.2 experiments with another measure of paper quality/productivity—the order an article appeared in an issue. It has no noticeable impact on the coefficient of interest or its standard error.

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

[19] 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.

[20] 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.

[21] While the information I found was publicly available, I apologise for the obvious intrusion. Given its sensitive nature, children's birth years are not currently available on my website (unlike other data in this paper).

Table 2: Readability scores

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

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

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.

2.1 Measuring readability

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

Although hundreds exist, I concentrate on the five most widely used, tested and reliable measures for adult reading material: Flesch Reading Ease, Flesch-Kincaid, Gunning Fog, SMOG (Simple Measure of Gobbledegook) and Dale-Chall (DuBay, 2004). Each are listed in Table 2.

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

The constants in each formula vary widely as do the components used to rank vocabulary. The Flesch Reading Ease and Flesch-Kincaid scales rely on syllable count, Gunning Fog and SMOG total polysyllabic words (words with three or more syllables) while Dale-Chall tallies words not on a pre-defined list of 3,000 so-called "easy" words.[22] These differences mean the four grade-level scores rarely generate identical figures; nevertheless, all five scores produce roughly equivalent rankings (Begeny and Greene, 2014).

Criticisms of readability scores are usually levied at their imprecision.[23] Evidence suggests they may not be accurate enough to adequately assess or guide development of legal briefs (Sirico, 2007), financial disclosure documents (Loughran and Mcdonald, 2014) or school reading material (Ardoin et al., 2005; Powell-Smith and Bradley-Klug, 2001). But despite poor accuracy, readability scores *do* correlate with reading difficulty (Begeny and Greene, 2014; DuBay, 2004; Francis et al., 2008; Hintze and Christ, 2004) making them appropriate measures to estimate gender differences in large samples.[24]

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

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

[24] At a bare minimum, no study (to my knowledge) has ever shown that any of the five scores used here are significantly inversely related to reading difficulty. Evidence from Begeny and Greene (2014) suggests the four grade-level readability scores, and particularly the SMOG and Dale-Chall scores, are more accurate for higher

A second criticism of readability scores is practical. Some programs that calculate them rely on unclear, inconsistent and possibly inaccurate algorithms to count words, sentences and syllables and determine whether a word is on Dale-Chall's easy word list (for a discussion, see Sirico, 2007). Additionally, features of the text—particularly full stops used in abbreviations and decimals in numbers—frequently underestimate average words per sentence and syllables per word.[25]

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

To determine sentence count, the program replaces common abbreviations with their full text,[26] decimals with a zero and deletes question and exclamation marks used in an obvious, mid-sentence rhetorical manner.[27] 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.[28]

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 byTeX's typesetting system. `libhyphen` is employed by most open source text processing software, including OpenOffice.

3 Results

Table 3 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 4 presents coefficients from an ordinary least squares (OLS) regression of the ratio of female co-authors on the five readability scores. To account for error correlation by editorial policy, observations are grouped by journal editor/editorial board and standard errors are adjusted accordingly.[29]

ability readers. (The study did not assess the Flesch Reading Ease score.)

[25] Typesetting code used to render equations—common in *Econometrica* abstracts published before 1980—also affects the accuracy of readability scores. I therefore manually replaced all such code with equivalent unicode characters. When no exact replacement existed, characters were chosen that mimicked as much as possible the equation's original intent while maintaining the same character and word counts. Readability scores were determined using the modified text.

[26] 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.

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

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

[29] Standard errors are very similar when clustering at the volume-, issue- or paper-level (see Appendix D.1).

Table 3: Textual characteristics per sentence, by gender

	Men	Women	Difference
No. characters	134.74	130.21	4.53***
	(0.43)	(1.45)	(1.56)
No. words	24.16	23.05	1.11***
	(0.08)	(0.26)	(0.29)
No. syllables	40.66	38.64	2.02***
	(0.13)	(0.45)	(0.48)
No. polysyllabic words	4.69	4.31	0.39***
	(0.02)	(0.07)	(0.08)
No. difficult words	9.38	8.90	0.48***
	(0.03)	(0.12)	(0.13)

Notes. Sample 9,123. 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.

Table 4: Gender differences in readability, article-level analysis

	(1)	(2)	(3)	(4)	(5)	(6)
Flesch Reading Ease	0.91*	0.88*	0.85*	0.82*	0.97*	0.53
	(0.48)	(0.48)	(0.50)	(0.49)	(0.52)	(0.54)
Flesch-Kincaid	0.19*	0.18*	0.19*	0.19*	0.22*	0.23*
	(0.11)	(0.11)	(0.11)	(0.11)	(0.12)	(0.12)
Gunning Fog	0.34***	0.33***	0.34***	0.34***	0.37***	0.34**
	(0.12)	(0.12)	(0.12)	(0.13)	(0.14)	(0.14)
SMOG	0.21**	0.21**	0.22**	0.21**	0.23**	0.19*
	(0.09)	(0.09)	(0.09)	(0.09)	(0.10)	(0.10)
Dale-Chall	0.10**	0.10**	0.10**	0.10**	0.11**	0.10*
	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.05)
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓
Institution effects				✓	✓	✓
Productivity effects					✓	✓
JEL effects						✓

Notes. 9,123 articles in (1)–(5); 5,216 articles in (6). Figures estimate the effect of an article's female ratio from an OLS regression on each readability score. Standard errors clustered on editor in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Table 5: Journal readability, comparisons to *AER*

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

Notes. Figures are the estimated coefficients on the journal dummy variables from (2) in Table 4. 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.

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.[30] Percentage-wise, women write 1–2 percent better than men.[31]

Column (2) includes 63 year dummies; column (3) adds another 182 journal and year interaction dummies; columns (4) and (5) introduce the 64 institution and 30 quality/productivity dummies, respectively.[32] Including these controls has little effect. Coefficients and standard errors are very similar to those in the first column.

The coefficients on the journal dummies in (2) are presented in Table 5. 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.

Column (6) in Table 4 controls for primary *JEL* classification. Since only post–1990 *JEL* classifications are used, estimates in (6) exclude 40 percent of the data. Nevertheless, coefficients are roughly equivalent—with the exception of the Flesch Reading Ease score; it halves and loses significance.

Figure 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) and (iv) controls for editor, journal, year, institution and author quality/productivity.[33] Due to small sample sizes—particularly of female authors—Figure 1 includes 561 articles from *AER Papers & Proceedings*.[34] *AER Papers & Proceedings* is coded as a separate journal and edited by the American Economic Association's president-elect.

The pink vertical line in Figure 1's left-hand graph is the marginal effect of female au-

[30] 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.

[31] Quotient of the coefficient on female ratio divided by the effect for men (ratio of zero) estimated at other covariates' observed values (see Appendix C.1).

[32] Appendix D.2 controls for the order an article appears in an issue—another measure of a paper's quality. Results are similar to those in Table 4.

[33] 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.

[34] *AER Papers & Proceedings* does not publish abstracts in its print version; only select years and papers are available online (2003 and 2011–2015), all of which are included. Appendix D.3 reproduces Figure 1, excluding *Papers & Proceedings* articles.

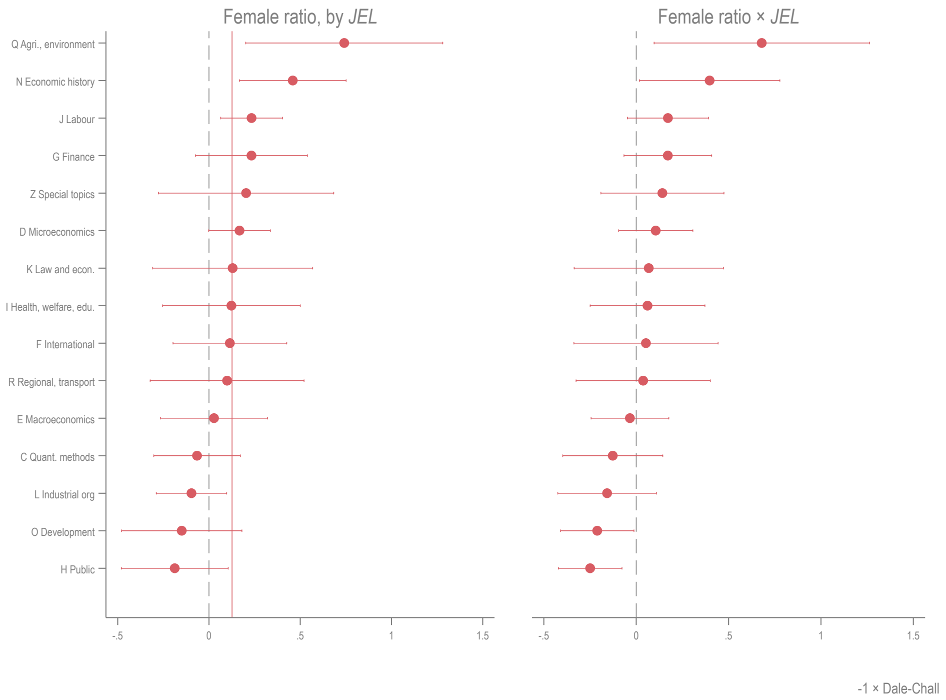

Figure 1: Gender differences in readability, by *JEL* classification

Notes. Sample 5,777 articles, including 561 from *AER Papers & Proceedings* (see Footnote 34). Codes A, B, M and P dropped due to small sample sizes of female-authored papers (see Footnote 33). Estimates from an OLS regression of:

$$-1 \times \text{Dale-Chall}_j = \beta_0 + \beta_1 \text{female ratio}_j + \boldsymbol{\beta}_2 \mathbf{J}_j + \boldsymbol{\beta}_3 \, \text{female ratio}_j \times \mathbf{J}_j + \boldsymbol{\theta} \, \mathbf{X}_j + \varepsilon_j \, ,$$

where 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 and quality/productivity dummies and ε_j is the error term. Left-hand graph shows marginal effects of female ratio for each *JEL* code ($\beta_1 + \beta_3^k$); the pink vertical line is the mean effect at observed *JEL* codes (0.125, standard error 0.046). Right-hand graph displays interaction terms (β_3^k). Horizontal lines represent 90 percent confidence intervals from standard errors adjusted for clustering on editor.

thorship at the mean. Its estimate coincides with results in Table 4—women's papers require six fewer weeks of schooling to understand—and is highly significant. Points reflect marginal effects across *JEL* classification (bars represent 90 percent confidence intervals from standard errors clustered by editor). Women earn higher marks for clarity in 11 out of 15 categories; only three are at least 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 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 1 suggests two things. First, the mostly insignificant interaction terms indicate outlier fields are probably not driving journals' gender readability gap—nor is any specific field bucking the trend. Second, the number of women in a field appears to have little effect on the size of the gap: Agriculture/Environment has one of the lowest concentrations of female-authored papers—but Economic History has one of the highest (Labour Economics falls between the two). Of course, Economic History papers are still overwhelmingly—as in 74 percent—penned just by men. But given the readability gap is present in subfields with both above- and below-average rates of sole female authorship, women may need to be better writers even where more of them publish.

In the remainder of the paper, I do not explicitly control for *JEL* classification given comparable codes are available for only a subset of the data and Table 4 and Figure 1 suggest they are relatively unimportant.

3.1 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 and 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):

$$\begin{aligned} \text{score}^s_{j_{it}} = &\beta_0 \, \text{score}^s_{it-1} + \beta_1 \, \text{female ratio}_j + \beta_2 \, \text{female ratio}_j \times \text{male}_i \\ &+ \beta_3 \, N_j + \boldsymbol{\theta} \, \mathbf{X}_j + \alpha_i + \varepsilon_{it}. \end{aligned} \tag{1}$$

$\text{score}^s_{j_{it}}$ is readability score s for article j—the article corresponding to author i's tth publication; score_{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 (0 for women, 1 for men) and the sex of his co-authors, respectively. N_j controls for author i's proportional contribution to paper j. $\mathbf{X}_j$ is a vector of editor, journal, year, journal × year and institution dummies,[35] 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 $1/N_j$.[36] Standard errors are adjusted for two-way clustering on editor and author.

Table 6 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.[37] 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 6 suggests they don't. While the interaction terms (β_2) are insignificant—*i.e.*, the observed disparity is plausibly due to chance[38]—the difference may reveal an increasing, convex relationship between female ratio and readability. Men's smaller effect may reflect their disproportionate tendency to co-author exclusively with other men—precisely where the marginal impact of an additional woman is low.[39]

[35] To remain consistent with a similar regression in Table 11, author productivity controls are omitted. Including them has almost no impact on estimates or their standard errors (see Appendix D.4).

[36] Assigning equal weight to all observations results in quantitatively and qualitatively similar results (see Appendix D.5).

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

[38] Specifically, the "chance" an author is included in the estimation sample: authors with fewer than three publications are dropped due to inclusion of the lag and first-differencing.

[39] 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 lease one women—themselves; the average female ratio of their papers is 0.6 (0.46 and 0.53 excluding articles written entirely by women and solo-authored papers, respectively).

Table 6: Gender differences in readability, author-level analysis

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Female ratio (women)	2.33**	0.37**	0.66***	0.47**	0.23**
	(0.99)	(0.18)	(0.23)	(0.19)	(0.09)
Female ratio (men)	0.89	0.16	0.20	0.12	0.11
	(1.42)	(0.27)	(0.32)	(0.22)	(0.11)
Female ratio × male	−1.44	−0.21	−0.47	−0.35	−0.13
	(1.55)	(0.32)	(0.37)	(0.26)	(0.13)
N	0.10	0.03	0.04	0.00	0.02
	(0.22)	(0.05)	(0.06)	(0.04)	(0.02)
Lagged score	0.03*	0.04**	0.03	0.02	0.02*
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)
z-test for no serial correlation					
Order 1	−20.51	−15.94	−16.82	−19.54	−21.80
Order 2	0.51	−0.22	0.16	0.39	−0.16
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal × Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓

Notes. Sample 9,166 observations (2,826 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). 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.

The coefficient on a paper's total number of authors is negligible and insignificant. Tests for serial correlation indicate no model misspecification.

Authors' past scores have little influence on current readability. Coefficients on the lagged dependant variables are small, suggesting readability is mostly determined contemporaneously—possibly during the revision process, discussed in Section 3.2. Nevertheless, their uniform positivity and weak significance for a majority indicate modest persistence and is investigated in Section 3.3.

3.2 Revisions

Table 4 established a gender readability gap for abstracts published in top economics journals. Table 6 suggests it primarily forms contemporaneously. A possible contemporaneous cause is peer review—specifically referee demands for more revisions by female authors.

I investigate this hypothesis in two steps. In Section 3.2.1 I establish a causal link with peer review by analysing papers' readability pre- and post-review. In Section 3.2.2, I determine the cost of more scrutiny—at least at *Econometrica*—female-authored articles take substantially longer to complete peer review.

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

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

Summary statistics. Table 7 compares textual characteristics between versions. Means in the first three columns are of majority male-authored papers (female ratio strictly below 50 percent); the final three columns are majority female-authored papers (female ratio at or above 50 percent).

Abstracts are considerably altered during peer review. Table 7's first panel displays raw counts. Draft abstracts are longer—more characters, words and sentences—and denser—more syllables, polysyllabic words and difficult words. The biggest changes are made to female-authored papers: figures in column six are 20–30 percent higher (in absolute value) than those in column three.

Peer review's impact on readability, however, is unclear. Readability scores are weighted averages of the ratios of (i) total word or "hard" word to sentence count and (ii) hard word to word count. Between working paper and published versions, (i) decreases and (ii) increases (Table 7, second panel).[40] (i) Peer review shortens sentences and reduces hard words per sentence: in male-authored papers, sentences are 5 percent shorter and contain 26 percent fewer polysyllabic words; in female-authored papers, they are 7 percent shorter and contain 30 percent fewer polysyllabic words. (ii) As a fraction of total word count, however, syllables, polysyllabic words and difficult words rise. To wit, hard word counts and total word count decline, but the latter by proportionately more; their ratios increase: between 1–3 percent for men and 1–2 percent for women.

According to the majority of scores, peer review improves readability (Table 7, third panel), a finding consistent with similar investigations at medical journals (Biddle and Aker, 1996; Hayden, 2008; Roberts and Nolen-Hoeksema, 1994).[41] Thanks to fewer hard words per sentence, SMOG scores are higher in published articles regardless of gender (see Table 2). In female-authored papers, the net effect for remaining scores is similarly positive. In male-authored papers, however, only the Gunning Fog and Flesch-Kincaid scores indicate a positive net effect; for the Flesch Reading Ease and Dale-Chall scores, it's negative. In any case, women's papers endure comparatively greater cuts in hard words relative to total words and larger falls in words per sentence; their abstracts always become more readable during peer review than do those by men.

Figure 2 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

[40] 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).

[41] 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, 2002b; 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 7: Textual characteristics, published papers vs. drafts

	Men			Women		
	Working paper	Published article	Difference	Working paper	Published article	Difference
No. sentences	6.50	5.10	−1.403***	6.77	5.06	−1.703***
	(0.06)	(0.04)	(0.054)	(0.15)	(0.08)	(0.139)
No. characters	864.90	649.23	−215.668***	908.61	635.75	−272.861***
	(7.15)	(4.63)	(7.119)	(18.50)	(10.29)	(18.413)
No. words	156.18	115.61	−40.575***	164.55	113.62	−50.927***
	(1.31)	(0.85)	(1.316)	(3.42)	(1.91)	(3.428)
No. syllables	257.76	193.23	−64.531***	269.41	187.67	−81.740***
	(2.13)	(1.39)	(2.123)	(5.53)	(3.08)	(5.495)
No. polysyllabic words	28.44	21.81	−6.624***	29.03	20.60	−8.429***
	(0.28)	(0.18)	(0.244)	(0.71)	(0.41)	(0.623)
No. difficult words	58.70	44.61	−14.085***	60.42	42.33	−18.088***
	(0.51)	(0.33)	(0.480)	(1.30)	(0.74)	(1.201)
No. words / sentence count	24.67	23.55	−1.117***	25.02	23.15	−1.869***
	(0.13)	(0.11)	(0.121)	(0.33)	(0.27)	(0.301)
No. polysyllabic words / sentence count	6.05	4.45	−1.593***	6.08	4.23	−1.850***
	(0.07)	(0.03)	(0.060)	(0.18)	(0.08)	(0.154)
No. syllables / word count	1.66	1.68	0.019***	1.64	1.66	0.012***
	(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	0.004**
	(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	0.005**
	(0.00)	(0.00)	(0.001)	(0.00)	(0.00)	(0.002)
Flesch Reading Ease	41.60	41.14	−0.453**	42.32	43.17	0.843*
	(0.26)	(0.18)	(0.181)	(0.66)	(0.41)	(0.429)
Flesch-Kincaid	−13.59	−13.37	0.214***	−13.57	−12.99	0.582***
	(0.06)	(0.05)	(0.048)	(0.15)	(0.11)	(0.120)
Gunning Fog	−17.24	−17.03	0.210***	−17.17	−16.57	0.598***
	(0.07)	(0.05)	(0.053)	(0.18)	(0.13)	(0.137)
SMOG	−15.12	−15.00	0.116***	−15.06	−14.69	0.368***
	(0.05)	(0.03)	(0.035)	(0.13)	(0.09)	(0.091)
Dale-Chall	−10.84	−10.93	−0.092***	−10.71	−10.70	0.019
	(0.02)	(0.02)	(0.015)	(0.06)	(0.04)	(0.036)

Notes. Sample 1,708 published articles authored by more than 50 percent men (1,735 NBER working papers); 270 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. Third and sixth columns subtract working paper figures (columns 1 and 4) from published article figures (columns 2 and 5) for men and women. Standard errors in parentheses. ***, ** and * difference statistically significant at 1%, 5% and 10%, respectively.

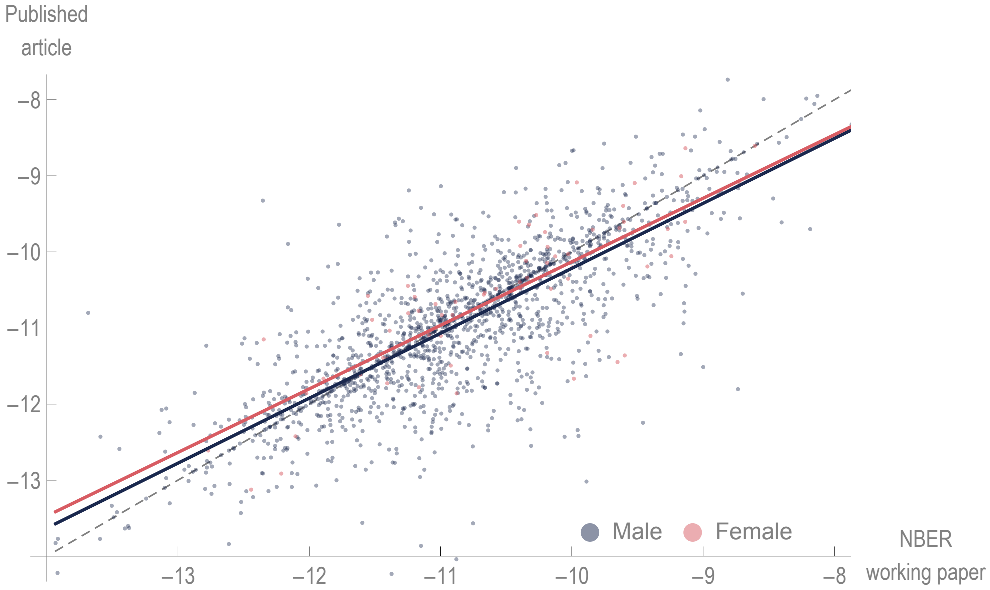

Figure 2: Published paper vs. draft readability

Notes. Sample 1,651 NBER working papers; 1,622 published articles. Data points represent each abstract's $-1 \times$ Dale-Chall score prepublicaction (NBER working paper) plotted against its $-1 \times$ Dale-Chall post-publication score. Pink represents women co-authoring only with other women (66 NBER working papers; 63 published articles); blue are men co-authoring only with other men (1,585 NBER working papers; 1,559 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.

line is a 45 degree line through the origin. As might be expected, poorly written draft abstracts emerge more readable in the published version (above the 45 degree line); abstracts that were already well written come out slightly less so (below the 45 degree line). Regardless, female-authored published papers are again more readable than they were as working papers relative to male-authored papers—further evidence that women's papers are more heavily scrutinised during peer review.[42]

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.

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

where score^s_{jP} and score^s_{jW} are readability scores s 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 and journal-year interaction dummies) and unobservable components, respectively. ε_{jP} is P's error term.

[42] An alternative hypothesis consistent with Figure 2 is that male-authored papers are scrutinised more, but edits made as a result reduce readability. The more substantial changes made to female-authored papers documented in Table 7, however, contradicts this theory.

P-specific fixed effects may be correlated with score^s_{jW}. Even if μ_{jP} and female ratio_j remain independent, positive correlation between score^s_{jW} and female ratio_j (Table 7) still biases OLS estimates of β_{1P} in a direction opposite to the bias on score^s_{jW}. Equation (3) eliminates the distortion by subtracting score^s_{jW} from both sides of Equation (2):

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

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

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

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

where β_{0W} is a constant specific to version W and β_{1W} reflects female ratio_j's impact on readability prior to peer review. $\mathbf{X}_{jW}$ and μ_{jW} are version-invariant observable (publication year and productivity dummies) and unobservable components, respectively.[43] ε_{jW} is version W's error term.

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

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

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

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

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 score_{jP}'s reduced form by substituting Equation (6) into Equation (2):

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

where $\widetilde{\boldsymbol{\theta}}_P = \boldsymbol{\theta}_P + \boldsymbol{\delta}_P$ and $\widetilde{\varepsilon}_{jP} = \widetilde{\varepsilon}_{jW} + \varepsilon_{jP}$. Equation (6) and Equation (7) are explicitly estimated via feasible GLS (FGLS). β_{1P} is identifiable post-estimation by subtracting reduced form coefficients; assuming zero partial correlation between μ_{jP} and female ratio_j, it is unbiased.[44]

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}.[45]

[43] 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. Because all papers in both samples share the same highest-ranked institution (NBER), institution effects are also omitted.

[44] 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).

[45] 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.

Table 8: The impact of gender, specific to peer review

	OLS	FGLS			OLS
	Published article	Working paper	Published article	Difference	Change in score
Flesch Reading Ease	1.75**	1.85*	3.31***	1.45**	1.33*
	(0.83)	(0.98)	(1.17)	(0.63)	(0.75)
Flesch-Kincaid	0.50**	0.17	0.63**	0.46***	0.44**
	(0.21)	(0.23)	(0.25)	(0.18)	(0.21)
Gunning Fog	0.53**	0.27	0.73***	0.46**	0.43*
	(0.21)	(0.24)	(0.26)	(0.18)	(0.24)
SMOG	0.32**	0.20	0.49***	0.28**	0.26*
	(0.14)	(0.16)	(0.17)	(0.12)	(0.15)
Dale-Chall	0.19***	0.31***	0.46***	0.15***	0.15**
	(0.05)	(0.11)	(0.10)	(0.04)	(0.06)
Editor effects	✓	✓	✓		✓
Journal effects	✓	✓	✓		✓
Year effects	✓	✓	✓		
Journal×Year effects	✓	✓	✓		✓
Productivity effects	✓	✓	✓		

Notes. Sample 2,008 NBER working papers; 1,978 published articles. Column one displays coefficients on female ratio (β_{1P}) from estimating Equation (2) directly via OLS (see Appendix C.3 for coefficients on score$^{s}_{jW}$); 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); robust standard errors in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Roughly restated, non-peer review factors must be either independent of its timing (and therefore subsumed in version-invariant fixed effects) or unrelated to gender.[46] Section 4.1 evaluates this assumption; briefly, however, I could think of nothing that simultaneously (and convincingly) influences readability, coincides with peer review's timing and correlates with author gender.[47]

Results. Table 8 presents results. They strongly indicate the readability gap grew precisely while papers were being reviewed. The first column displays β_{1P} from OLS estimation of Equation (2). According to all five scores, women's readability gains outpace men's between versions. Estimates additionally confirm published readability is correlated with draft readability: coefficients on score$_{jW}$ (shown in Appendix C.3) are positive and significant—but only about 0.8. A less than unit value suggests μ_{jP} exerts downward pressure on score$_{jW}$'s coefficient, thereby artificially inflating first column figures (see previous section).

Table 8's remaining columns present results from both strategies meant to deal with this bias. Columns 2–4 display FGLS estimates. Coefficients on female ratio$_j$ from Equation (6) ($\widetilde{\beta}_{1W}$) and Equation (7) ($\widetilde{\beta}_{1W} + \beta_{1P}$) are shown in columns two and three, respectively.

[46] 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$_j$.

[47] A possible exception is external feedback solicited outside of peer review—*e.g.*, during conferences and seminars. As Section 4.1 points out, however, the population of people who provide such feedback overlaps with the population of journal referees. It seems unlikely that this population is biased only in one setting—especially given both settings emphasise gender neutrality.

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 60–70 percent of the total (biased) effect of female ratio in Equation (7); Flesch Reading Ease and Dale-Chall scores indicate a smaller proportion (30–40 percent).[48] Column four displays their difference (β_{1P}); it is positive and significant or highly significant for all five scores.

OLS estimates of β_{1P} from Equation (3) are shown in Table 8's final column. Magnitudes are close to FGLS estimates—confirming earlier conclusions—standard errors are slightly higher. Both strategies show a significant increase in the gender readability gap *ex post*; assuming non-peer review factors are always independent of either its timing or gender, this establishes the desired causal link.[49]

3.2.2 Duration of peer review. Women write better than men. So what? If writing well takes no more time than writing poorly then, indeed, so what. But both gut feeling and hard evidence suggest this can't be the case. "Writing simply and directly only looks easy" (Kimble, 1994, p. 53). Revising, redrafting and selecting just the right word is hard work. 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).

So if writing simply and directly takes time—and assuming referee demands are a source of journals' gender readability gap—one observable repercussion will be prolonged peer review for female authors. To investigate, I turn to *Econometrica*, the only journal to make disaggregated data on the revision process publicly available.

Figure 3 is a histogram of time (in months) between dates papers are first submitted to and their final revisions received by *Econometrica*'s editorial office. Blue bars represent articles written only by men, pink bars are those just by women; papers co-authored by men and women are not included.

Since 1950, *Econometrica* published 53 papers authored entirely by women.[50] As Figure 3 illustrates, their review times disproportionately cluster in the distribution's right tail: articles by women are six times more likely to experience delays above the 75th percentile than they are to enjoy speedy revisions below the 25th. In fact—and despite making up just 2 percent of the sample—one such paper holds the record for longest review: Andrea Wilson's "Bounded Memory and Biases in Information Processing" (November, 2014). Ms. Wilson's paper took a decade to get published.

For a more precise appraisal, I build on a model by Ellison (2002b, Table 6, p. 963) and estimate Equation (8):

$$\begin{aligned}\text{revision duration}_j = {} & \beta_0 + \beta_1 \,\text{female ratio}_j + \beta_2 \,\text{mother}_j + \beta_3 \,\text{birth}_j \\ & + \beta_4 \,\max t_j + \beta_5 \,\text{no. pages}_j + \beta_6 \,N_j \\ & + \beta_7 \,\text{order}_j + \boldsymbol{\theta}\,\mathbf{X}_j + \varepsilon_j,\end{aligned} \tag{8}$$

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

[49] The wider gap post-peer review suggests causality with peer review; it does not establish causality with referee scrutiny. This issue and the assumption of timing-independence are discussed in Section 4.1; the latter is summarised briefly in the previous section. The discussion in Footnote 46 also applies to the precise accuracy of the assumption's phrasing used here.

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

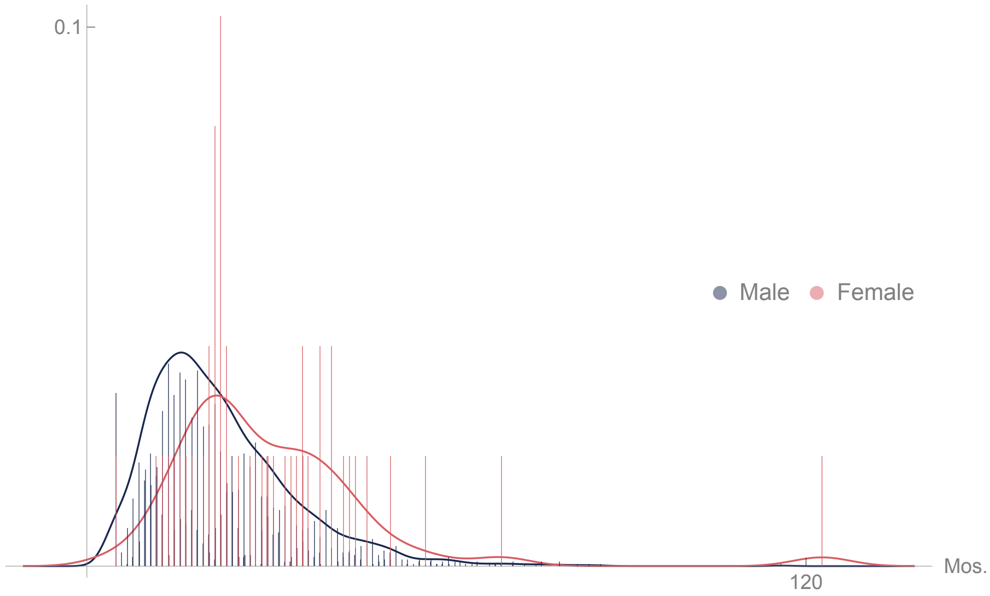

Figure 3: Distribution of review times at *Econometrica*

Notes. Sample 2,445 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 (2,396); pink bars are papers written only by women (49). Source: *Econometrica*.

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,[51] $\max t_j$ 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 and order_j is the order in which article j appeared in an issue.[52]

Table 9 displays results. Every paper published in *Econometrica* undergoes extensive review, but the large and highly significant coefficient on female ratio suggests women bear the worst of it.[53] The average male-authored paper takes 18.5 months to complete all revisions; papers by women need half a year longer.[54]

Why? Well, it's not because of motherhood. Yes, giving birth slows down review—responding to referees is apparently put on hold for the first year of a newborn's life—but having a young child has the opposite effect. A pause for childbirth is expected; a productivity boost from pre-schoolers is not. Perhaps wanting to spend time with the kids motivates women to get organised? Or, maybe the most organised women are the only ones having children? The former suggests motherhood is not the productivity killer it's rumoured to be—at least among highly educated women. The latter implies only superstar women feel academic careers and motherhood are simultaneously manageable.[55] Both interpretations are provocative, but should be made with caution given (i) counter-intuitive results, (ii) obtaining an unbiased es-

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

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

[53] This conclusion is robust to altering the age-threshold on mother_j (see Appendix D.6).

[54] Male effect estimated with zero female co-authors (standard error 0.228).

[55] 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

TABLE 9: Revision duration at *Econometrica*

	Time (in months)	Standard error
Female ratio	6.030***	2.089
Mother	−13.004***	3.674
Birth	10.244**	4.650
Max t	−0.132***	0.033
No. pages	0.182***	0.025
N	0.970**	0.393
Order	0.229***	0.070
Constant	37.389***	1.798
Editor effects	✓	
Year effects	✓	
Institution effects	✓	

Notes. Sample 2,626 articles. Coefficients from OLS estimation of Equation (8) with robust standard errors. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

timate of β_2 was *not* this study's objective and (iii) mother$_j$ equals one for only 13 articles in the sample.[56]

As for Table 9's remaining coefficients, all are significant or highly significant and correspond to earlier estimates by Ellison (2002b). 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 enjoy marginally faster reviews.[57]

3.3 Experience

Does the readability gap change as publication counts increase? Yes, it widens—from women writing more clearly and men possibly less so. As their careers evolve, women improve: their average readability scores are 1–5 percent higher than the readability of their first papers; their latest papers 1–7 percent higher (Appendix B.1). For a man, however, his average and last paper may be more poorly written than the first. Figure 4 plots mean Flesch Reading Ease scores grouped by authors' tth article; as the count increases, men and women diverge.

Table 10 tests significance of that divergence by FGLS estimation of Equation (1) (omitting score$^s_{it-1}$) on subsamples corresponding to authors' first ($t = 1$), second ($t = 2$), third ($t = 3$), fourth and fifth ($t = 4$–5) and sixth and up ($t \geq 6$) articles published in the journals and time periods covered by the data. Only marginal effects on co-authoring with women for female authors are shown (β_1). Final column is a population-averaged estimate on the pooled sample. Regressions in columns ($t = 1$) to ($t \geq 6$) are weighted by $1/N_j$ (see Section 3.1), 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 despite falling numbers of observations.[58] Between columns

remarkably difficult to find out—I (perhaps unfairly) give this interpretation less weight.

[56] The count increases to 14 and 16 articles when mother$_j$'s threshold is defined as children younger than ten and 18, respectively (see Appendix D.6).

[57] Ellison (2002b)'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.

[58] Coefficient equality test statistics are available in Appendix C.4.

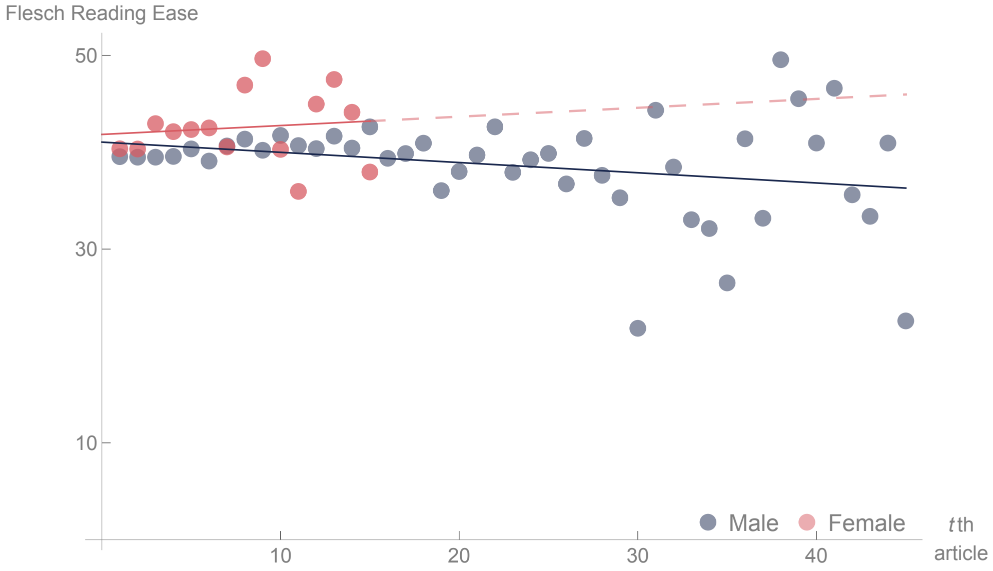

Figure 4: Readability of authors' tth publication

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).

Table 10: Gender gap in readability at increasing t

	$t = 1$	$t = 2$	$t = 3$	$t =$ 4–5	$t \geq 6$	All
Flesch Reading Ease	0.38	1.45*	4.53***	2.97	3.09	1.57**
	(0.61)	(0.81)	(1.00)	(1.85)	(2.29)	(0.74)
Flesch-Kincaid	0.10	0.16	0.74***	0.55	0.73*	0.19
	(0.14)	(0.21)	(0.22)	(0.39)	(0.41)	(0.15)
Gunning Fog	0.23	0.37	1.02***	0.78*	0.97*	0.41**
	(0.16)	(0.25)	(0.27)	(0.44)	(0.51)	(0.19)
SMOG	0.14	0.25	0.70***	0.61*	0.66*	0.31**
	(0.12)	(0.17)	(0.19)	(0.35)	(0.37)	(0.13)
Dale-Chall	0.08	0.08	0.37***	0.28*	0.41*	0.18**
	(0.06)	(0.07)	(0.12)	(0.16)	(0.24)	(0.07)
No. observations	6,900	2,826	1,674	1,901	2,765	11,992
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓
Journal × Year effects						✓
Institution effects	✓	✓	✓	✓	✓	✓
Productivity effects						✓

Notes. β_1 from FGLS estimation of Equation (1) without lagged dependent variable. First column restricts sample to authors' first publication in the data ($t = 1$), second column to their second ($t = 2$), *etc.* Regressions weighted by $1/N_j$ (see Section 3.1). 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. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Table 11: Slope of t, by gender

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
t (men)	−0.01	−0.02	−0.02	−0.01	0.00
	(0.06)	(0.01)	(0.01)	(0.01)	(0.00)
t (women)	0.79**	0.17**	0.24***	0.17**	0.07
	(0.39)	(0.08)	(0.08)	(0.07)	(0.05)
Difference	0.80**	0.19**	0.25***	0.18**	0.07
	(0.40)	(0.08)	(0.09)	(0.07)	(0.05)
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓

Notes. Sample 9,166 observations, 2,826 distinct authors. First-differenced, IV estimation of Equation (9) (Arellano and Bover, 1995; Blundell and Bond, 1998). Figures for women: marginal effect of t for women co-authoring only with women ($\beta_3 + \beta_5$); figures for men: marginal effect of t for men co-authoring only with men ($\beta_3 + \beta_4$). 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.

($t = 1$) and ($t = 2$), the gap marginally widens but is not significant; after that, it triples (at least); the increase is significant ($p < 0.05$) for all five scores.[59] At higher publication counts, differences in male-female readability remain roughly constant, although estimates are only weakly significant and suffer from small samples of female authors.

Despite having the largest sample, first-time publications are not driving the observed readability gap. Figure 4 suggests little or no gender difference when $t = 1$; Table 10 backs this up. Coefficients in column ($t = 1$) are imprecise, roughly half the size of those from a pooled regression (last column) and a fraction the size of estimates in columns ($t = 3$), ($t =$ 4–5) and ($t \geq 6$). Wald tests (Appendix C.4) reject equality of β_1 in the first and third models at $p < 0.01$ for the Flesch Reading Ease, Flesch-Kincaid and SMOG scores and $p < 0.05$ for the Gunning Fog and Dale-Chall scores.

To generate a slope coefficient measuring the mean additional contribution each paper makes to readability, I estimate:

$$\begin{aligned} \text{score}^s_{jit} &= \beta_0\,\text{score}^s_{it-1} + \beta_1\,\text{female ratio}_j + \beta_2\,\text{female ratio}_j \times \text{male}_i \\ &\quad + \beta_3\,t + \beta_4\,t \times \text{male}_i + \beta_5\,\text{female ratio}_j \times t \\ &\quad + \beta_6\,\text{female ratio}_j \times t \times \text{male}_i + \beta_7\,N_j + \boldsymbol{\theta}\,\mathbf{X}_j + \alpha_i + \varepsilon_{it}. \end{aligned} \tag{9}$$

Equation (9) includes t and its interaction with an article's female ratio and the author's sex; otherwise, it is identical to Equation (1) and similarly estimated by first-differencing with endogeneity instrumented with earlier lags (Arellano and Bover, 1995; Blundell and Bond, 1998).

t's (immediate) marginal effect is presented in Table 11. Male effects estimated for male authors co-authoring with no females ($\beta_3 + \beta_4$); female effects for female authors co-authoring with no males ($\beta_3 + \beta_5$). The coefficients agree: readability scores remain constant as men publish more papers—all are very close to zero and none significantly different from it. For women, however, every additional paper is more readable than the last; the effect is statistically significant for four out of five scores.

[59] Note that figures in columns two and three of Table 10 are roughly in line with third column estimates in Table 8—on average, $t = 2.7$ for female-authored articles released first as NBER working papers.

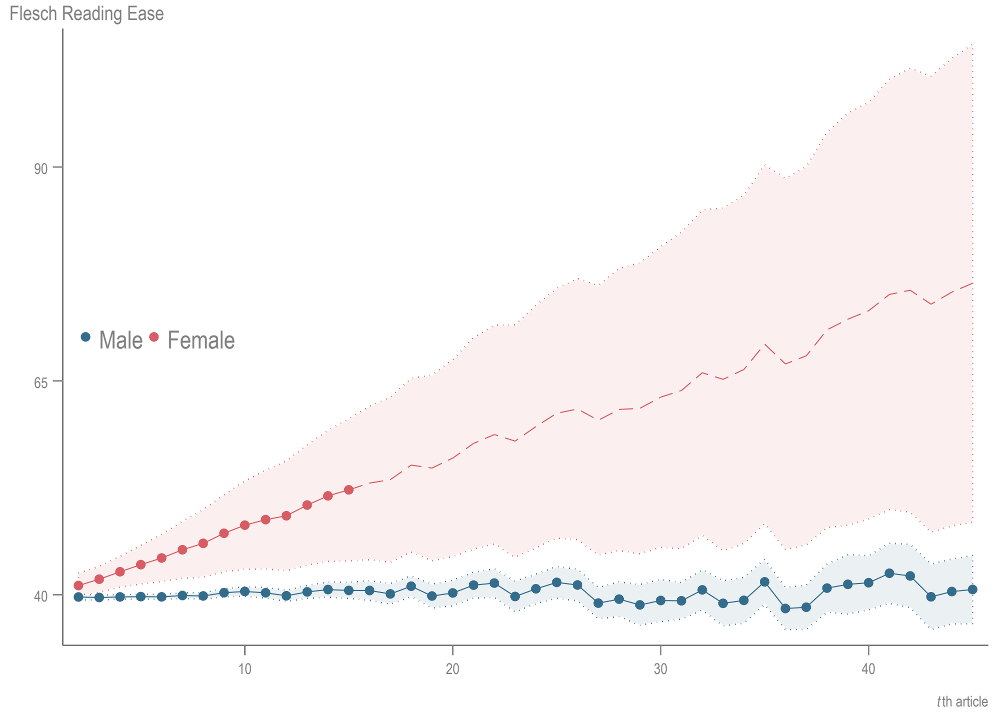

FIGURE 5: Readability of authors' tth publication

Notes. Flesch Reading Ease marginal mean scores for authors' first, second, ..., tth, ... publication in the data. Estimates derived from first-differenced IV estimation of Equation (9) (Arellano and Bover, 1995; Blundell and Bond, 1998) presented in Table 11. Pink represents women co-authoring only with other women; blue are men co-authoring only with other men. Dotted line indicates out-of-sample forecast.

Figure 5 plots mean male and female effects over t using estimates derived from the Flesch Reading Ease regression in Table 11. It serves as a more precise illustration of the trends depicted in Figure 4. As before, there is little or no gender difference in readability for the first two publications but thanks to women's self-improvement, the gap widens substantially after that. While Figure 4 suggested men were also writing more poorly, that effect is not present here.

3.3.1 An explanation? Table 10 documents a rise in the readability gap as women publish more articles. Table 11 points to an individual-specific explanation driving this phenomenon. One such explanation is "learning-by-doing". If the payoff from lucid exposition is high, people will catch on—either by internalising explicit comments on text readability in referee reports from earlier papers or making the (un)conscious connection that review times are faster when text is clearer.[60] Applying that payoff only to women yields a succinct explanation for the gap's observed growth.

But there are two possible alternatives. I investigate both and show that neither is likely. As discussed in Section 3.1, readability and female ratio may be nonlinearly related—specifically, evidence suggests the latter is increasing and convex in the former. Thus, if women are more likely to co-author with other women (or by themselves) when they already have several publications behind them, the observed increase would actually reflect that late-career concentration.

[60] A related possibility is that women are more responsive to referee reports. This and other explanations are addressed in Section 4.1.

Table 12: Impact of t on female ratio

	Estimate	Standard error
t (men)	0.00160**	0.00074
t (women)	−0.03968***	0.00761
Editor effects	✓	
Journal effects	✓	
Year effects	✓	
Journal×Year effects	✓	
Institution effects	✓	
Productivity effects	✓	

Notes. Sample 11,992 observations (2,826 authors). Figures from fixed effects estimation of

$$\text{female ratio}_{j_{it}} = \beta_1\, t + \beta_2\, t \times \text{male}_i + \boldsymbol{\theta}\, \mathbf{X}_j + \alpha_i + \varepsilon_{it}.$$

t (men): marginal effect of t for male authors ($\beta_1 + \beta_2$); t (women): marginal effect of t for female authors (β_1). Regression 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.

This is not the case. Table 12 displays the marginal effect of t for men and women from a fixed effects regression on female ratio. The effect for men is positive; the effect for women negative. As t increases, genders diversify: men publish with more women and women publish with more men.

The second explanation applies to the evolution of female writers as a group—perhaps the clearest women publish more often? Equation (10) investigates:

$$\begin{aligned}\Phi\left(T_i \geq x\right) = \Phi\,(&\beta_0 + \beta_1 \text{ female ratio}_{i1} + \beta_2 \text{ score}^s_{i1} \\ &+ \beta_3 \text{ male}_i + \beta_4 \text{ female ratio}_{i1} \times \text{score}^s_{i1} \\ &+ \beta_5 \text{ female ratio}_{i1} \times \text{male}_i + \beta_6 \text{ score}^s_{i1} \times \text{male}_i \\ &+ \beta_7 \text{ female ratio}_{i1} \times \text{score}^s_{i1} \times \text{male}_i + \boldsymbol{\theta}\, \mathbf{X}_{i1} + \varepsilon_i)\,,\end{aligned} \tag{10}$$

where Φ is the standard normal cumulative distribution function, $T_i \geq x$ equals one if author i's total publication count, T_i, is at least the positive integer x and subscript 1 refers to his first publication. Table 13 displays $\Phi' \cdot (\beta_2 + \beta_4)$—the marginal effect of female authors' first paper readability scores on the probability of publishing multiple times.

The readability of a woman's first paper does not predict how often she will publish. Regardless of the threshold chosen, it has no impact on her eventual productivity and is indistinguishable from zero in all estimates.

4 Discussion

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

The gender readability gap grows between first draft and final publication and over the course of women's careers—implying it's due neither to article- nor author-specific fixed effects, exclusively. This rules out inborn advantages and one-off improvements in response to external circumstances unrelated to peer review.

Table 13: Effect of female first score on number of publications

x	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
3	0.00	-0.01	-0.01	0.00	0.04
	(0.00)	(0.01)	(0.01)	(0.01)	(0.04)
4	0.00	0.00	0.00	0.01	0.01
	(0.00)	(0.01)	(0.01)	(0.01)	(0.03)
5	0.00	-0.01	0.00	0.00	0.01
	(0.00)	(0.01)	(0.01)	(0.01)	(0.03)
Editor effect	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓

Notes. Sample 6,899 authors. Marginal effect of a female author's first paper readability score on the probability of publishing x or more times ($\Phi' \cdot (\beta_2 + \beta_4)$) from probit estimation of Equation (10). Standard errors clustered on editor in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Yet a variety of prospects are still feasible. I have frequently implied—and sometimes outright proposed—one possibility: bias among journal referees. This single explanation neatly accounts for all observed patterns. If referees apply higher standards to female-authored papers, those papers undergo more thorough review. Added scrutiny should improve women's exposition but lengthen review times—as seen in Section 3.2.2. The rewards from clearer writing are presumably internalised, meaning women gradually improve—which they do, as illustrated in Section 3.3.

Although no one (to my knowledge) has tested whether men and women receive different critical feedback in peer review reports, several studies document it in 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 and guidance or praise they are offered is vague (Correll and Simard, 2016).[61]

Students display a similar bias. Data from Rate My Professors suggest female lecturers should be "helpful", "clear", "organised" and "friendly". Men, instead, are praised (and criticised) for being "smart", "humble" or "cool" (Schmidt, 2015).[62] 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, 2015).

4.1 Alternative hypotheses?

While a concise explanation of observed trends cannot entirely abstract from peer review, there are alternative hypotheses that do absolve journal referees. One option turns from production of feedback toward its perception. Possibly, women are more sensitive to content in referee reports and receptive to its informational value. A different interpretation that leads to different

[61] 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).

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

action—*e.g.*, greater likelihood of making changes—would trigger the feminine inclination to write more clearly.[63]

Greater female responsiveness explains well article-level dynamics and differences in review time. It struggles to account for the gradual readability improvements women make over their lifetimes. For both phenomena to hold requires systematic failure among women to efficiently allocate resources: diligently addressing every referee concern has no apparent upside—acceptance rates are unaffected—and a very clear downside—constant redrafting takes time. Wouldn't rational women re-examine initial beliefs… and then start acting like men? Yet, this is not what we observe. Instead, 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.[64]

A second alternative implicates referee assignment. If women's papers are more likely assigned female referees and female referees are more demanding critics, clearer writing could reflect their tougher reviews.

This idea accounts for article- and author-level dynamics as well as women's longer review times. And of course, women concentrate in particular fields, so it's natural their papers are more often assigned female referees. However, for the readability gap to exist only because of specialisation, controlling for *JEL* classification should explain it.[65] It does not. In fact, even including 718 tertiary *JEL* category dummies has virtually no effect: the gap remains almost identical to (but standard errors somewhat higher than) estimates in Table 4, column (6).[66] So if referee assignment is causing the gap, it's only because journals disproportionately refer female-authored papers to the toughest critics. Meaning it isn't referees who are biased—it's editors.[67]

Some evidence supports this idea—female referees do indeed review more female-authored papers (Abrevaya and Hamermesh, 2012; Gilbert et al., 1994).[68] It's not so clear whether their reports are any more critical. A study specific to post-graduate biologists suggests yes (Borsuk et al., 2009); another analysing past reviews in an economics field journal does not (Abrevaya and Hamermesh, 2012).

A third possibility invalidates timing-independence, an assumption invoked in Section 3.2.1 to causally link the readability gap with peer review. One external factor in particular may co-

[63] 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).

[64] A related argument invokes an upside—acceptance rates increase—but failure by men to update beliefs.

[65] 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.

[66] Estimation mirrors Table 4, column (6) but 20 primary *JEL* category dummies are replaced with 718 tertiary ones. Coefficients on female ratio are as follows (standard errors in parentheses). Flesch Reading Ease: 0.49 (0.75); Flesch-Kincaid: 0.23 (0.15); Gunning Fog: 0.35 (0.18); SMOG: 0.21 (0.14); Dale-Chall: 0.09 (0.07).

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

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

incide with this timing: feedback women receive in conferences and seminars. Perhaps women tighten prose (before or after submission) in response to audience member remarks? Anecdotal evidence suggests female speakers are given a harder time,[69] although I could find no scientific analysis to support (or contradict) this claim.[70] Nevertheless, most participants are also current (or future) journal referees. Neutral review feedback is inconsistent with non-neutral presentation feedback when originating from the same group.[71]

A final alternative is rather uncomfortable. Perhaps female-authored manuscripts deserve more criticism because they are not as good? As mentioned earlier, factors correlated with gender but actually related to competency should decline when appropriate proxies are included. The sample itself is one such proxy—these are, after all, only articles published in the top four economics journals. Adding other controls—author institution, total article count and published order in an issue—has no impact.[72] 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. And if this is true, referees *should* peruse our papers more carefully—a byproduct of which could be better written papers after-the-fact or more attractive prose compensating for structural weaknesses before it.[73]

"Quality" is subjective; measurement is not easy. Nevertheless, attempts using citation counts and journal acceptance rates do not indicate that men's research is any better: as discussed in Section 1, gender has virtually zero impact on the latter;[74] 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).

4.2 Open review

More complicated, multi-factor explanations could resolve inconsistencies present when each is analysed in isolation. Perhaps female economists are both innately better writers and mostly non-native English speakers?[75] Maybe women are perfectionists, and it gets stronger with age?[76] And it is always possible that the statistically significant relationships this paper documents are nothing more than (unfortunate—for me) flukes.

[69] A related theory is that women receive more critical feedback in conferences and seminars because they present their work more often. In a survey of economists, Sarsons (2015) finds that men and women are equally likely to present co-authored work but women are actually *less* likely to present solo-authored work.

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

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

[72] Published order in an issue refers to the order an article appears in a particular issue (*i.e.*, one for the lead article, two for the second article, *etc.*). This control was introduced as a a set of indicator variables. See Appendix D.2 for regression output.

[73] It does seem contradictory, however, that women would be capable of writing better than men—even before referee input (Table 8)—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.

[74] 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.

[75] Innately better writing offsets poorer scores in initial papers; the rise in readability over time reflects improving English. A primitive surname analysis (Appendix A) does not, however, suggest non-native economists are more likely to be female.

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

Still, no explanation matches the simplicity and believability of biased referees. Coherence and economy do not establish fact, but they are useful guides, especially with no convincing alternative and a serious problem: assuming *Econometrica* is indicative, women spend too much time in peer 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" is much less paradoxical.[77]

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.[78] 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.[79] 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., 1999; van Rooyen et al., 2010). 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)[80]—but if more accountability promotes fairer outcomes, ethical arguments in its favour should outweigh minor practical concerns.

5 Conclusion

This paper makes a curious discovery: female-authored abstracts published in the top four economics journals are better written. I explore many possible reasons; the most straightforward is that referees apply higher standards to female-authored papers.

No prior study has uncovered convincing evidence of gender bias in journal acceptance rates. While it's encouraging that sex is irrelevant to publication outcomes, that does not mean it has no effect on the process. If female authors endure unfair criticism in referee reports, their clearer writing and longer review times are natural. Because review times are six months longer—at least at *Econometrica*—lower research output is too.

[77] 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).

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

[79] 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.

[80] 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).

The topic of my study is narrow, but its methodology has wider applications. To the best of my knowledge, it is the first to uncover subtle group differences with readability scores.[81] These scores are not new—all are extensively tested with well-documented properties—but their use is mostly confined to determining whether text is appropriate for intended audiences.[82] As this paper demonstrates, however, readability scores are also effective tools to evaluate asymmetry anywhere ideas are communicated orally or in writing and large amounts of source material are easily obtainable: journalism, speeches, student essays, business plans, Kickstarter campaigns, *etc.* Research potential is substantial.

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. DOI: 10.1162/REST_a_00163.

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

Ardoin, S. P. et al. (2005). "Accuracy of Readability Estimates' Predictions of CBM Performance." *School Psychology Quarterly* 20(1), pp. 1–22. DOI: 10.1521/scpq.20.1.1.64193.

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. DOI: 10.2307/2297968.

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. DOI: 10.1016/0304-4076(94)01642-D.

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.

Baker, K. J. (2015). *Should Academic Conferences Have Codes of Conduct?* Washington, DC.

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

Begeny, J. C. and D. J. Greene (2014). "Can Readability Formulas Be Used to Successfully Gauge Difficulty of Reading Materials?" *Psychology in the Schools* 51(2), pp. 198–215. DOI: 10.1002/pits.21740.

Berk, J. B., C. R. Harvey and D. Hirshleifer (2015). "Preparing a Referee Report: Guidelines and Perspectives".

Bernstein, R. (2015). "Updated: Sexist Peer Review Elicits Furious Twitter Response, PLOS Apology". *Science*, pp. 1–14. DOI: 10.1126/science.aab2568.

Biddle, C. and J. Aker (1996). "How Does the Peer Review Process Influence AANA Journal Article Readability?" *AANA journal* 64(1), pp. 65–8.

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

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

[82] Long and Christensen (2011), Lehavy et al. (2011) and Thörnqvist (2015) use readability scores in interesting, non-conventional ways. The former investigates whether a legal brief's Flesch Reading Ease score is correlated with its success on appeal (it is not); the latter two use readability measures to proxy for complex information in financial reports, finding less readable material is less informative (Lehavy et al., 2011), especially for non-sophisticated investors (Thörnqvist, 2015).

Blundell, R. and S. Bond (1998). “Initial Conditions and Moment Restrictions in Dynamic Panel Data Models”. *Journal of Econometrics* 87(1), pp. 115–143. DOI: 10.1016/S0304-4076(98)00009-8.

Boring, A. (2015). “Gender Biases in Student Evaluation of Teachers”. Paris, France.

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. DOI: 10.1525/bio.2009.59.11.10.

Budden, A. E. et al. (2008a). “Double-blind Review Favours Increased Representation of Female Authors”. *Trends in Ecology and Evolution* 23(1), pp. 4–6. DOI: 10.1016/j.tree.2007.07.008.

Budden, A. E. et al. (2008b). “Response to Webb et al.: Double-blind Review: Accept with Minor Revisions”. *Trends in Ecology and Evolution* 23(7), pp. 353–354. DOI: 10.1016/j.tree.2008.04.001.

Budden, A. E. et al. (2008c). “Response to Whittaker: Challenges in Testing for Gender Bias”. *Trends in Ecology and Evolution* 23(9), pp. 480–481. DOI: 10.1016/j.tree.2008.06.004.

Ceci, S. J. et al. (2014). “Women in Academic Science : A Changing Landscape”. *Psychological Science in the Public Interest* 15(3), pp. 75–141. DOI: 10.1177/1529100614541236.

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

Correll, S. and C. Simard (2016). “Research: Vague Feedback Is Holding Women Back”. *Harvard Business Review*.

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. DOI: 10.1037/0022-3514.81.2.322.

Dale, E. and J. S. Chall (1948). “A Formula for Predicting Readability”. *Educational Research Bulletin* 27(1), pp. 11–20.

DuBay, W. H. (2004). *The Principles of Readability*. Impact Information: Costa Mesa, California.

Ecklund, E. H. and A. E. Lincoln (2011). “Scientists Want More Children”. *PLoS ONE* 6(8), pp. 1–4. DOI: 10.1371/journal.pone.0022590.

Ellison, G. (2002a). “Evolving Standards for Academic Publishing: A q-r Theory”. *Journal of Political Economy* 110(5), pp. 994–1034. DOI: 10.1086/341871.

Ellison, G. (2002b). “The Slowdown of the Economics Publishing Process”. *Journal of Political Economy* 110(5), pp. 947–993. DOI: 10.1086/341868.

Faigley, L. and S. P. Witte (1981). “Analyzing Revision”. *College Composition and Communication* 32(4), pp. 400–414. DOI: 10.2307/356602.

Feingold, A. (1994). “Gender Differences in Personality: A Meta-analysis”. *Psychological Bulletin* 116(3), pp. 429–456. DOI: 10.1037/0033-2909.116.3.429.

Foschi, M. (1996). “Double Standards in the Evaluation of Men and Women”. *Social Psychology Quarterly* 59(3), pp. 237–254. DOI: 10.2307/2787021.

Francis, D. J. et al. (2008). “Form Effects on the Estimation of Students’ Oral Reading Fluency Using DIBELS”. *Journal of School Psychology* 46(3), pp. 315–342. DOI: 10.1016/j.jsp.2007.06.003.

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. DOI: 10.1257/jep.8.1.165.

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. DOI: 10.1001/jama.1994.03520020065018.

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. DOI: 10.1257/0895330042162386.

Goldberg, P. (1968). "Are Women Prejudiced against Women?" *Trans-action* 5(5), pp. 28–30. DOI: 10.1007/BF03180445.

Goldberg, P. K. (2015). "Report of the Editor: American Economic Review". *American Economic Review* 105(5), pp. 698–710. DOI: 10.1257/aer.15000011.

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. DOI: 10.1371/journal.pone.0148405.

Hamermesh, D. S. (1994). "Facts and Myths about Refereeing". *Journal of Economic Perspectives* 8(1), pp. 153–163. DOI: 10.1257/jep.8.1.153.

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. DOI: 10.1023/A:1025008802657.

Hartley, J., J. W. Pennebaker and C. Fox (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.

Hayden, J. D. (2008). "Readability in the British Journal of Surgery". *British Journal of Surgery* 95(1), pp. 119–124. DOI: 10.1002/bjs.5994.

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. DOI: 10.1037/0021-9010.90.5.905.

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

Hintze, J. M. and T. J. Christ (2004). "An Examination of Variability as a Function of Passage Variance in CBM Progress Monitoring". *School Psychology Review* 33(2), pp. 204–217.

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. DOI: 10.1111/1471-6402.00063.

Journal of Trauma and Acute Care Editorial Board (2015). *Journal of Trauma and Acute Care Surgery*. http://journals.lww.com/jtrauma/Documents/Editorial%20Board%20Reports/2015%200909%20AAST%20Ed%20Board%20print%20report.pdf. Accessed: 2016-10-04.

Kimble, J. (1994). "Answering the Critics of Plain Language". *Scribes Journal of Legal Writing* 51(1994-1995), pp. 51–85.

King, D. W., C. Tenopir and M. Clarke (2006). "Measuring Total Reading of Journal Articles". *D-Lib Magazine* 12(10), pp. 1082–9873.

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*. DOI: 10.1016/j.euroecorev.2016.02.017.

Kroll, B. (1990). “What Does Time Buy? ESL Student Performance on Home versus Class Compositions”. In: *Second Language Writing*. Ed. by B. Kroll. Cambridge University Press: Cambridge, U.K. Chap. 9, pp. 140–154. DOI: 10.1017/CBO9781139524551.014.

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. DOI: 10.2308/accr.00000043.

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

Lloyd, M. E. (1990). “Gender Factors in Reviewer Recommendations for Manuscript Publication”. *Journal of Applied Behavior Analysis* 23(4), pp. 539–543. DOI: 10.1901/jaba.1990.23-539.

Long, L. N. and W. F. Christensen (2011). “Does the Readability of Your Brief Affect Your Chance of Winning an Appeal?” *Journal of Appellate Practice and Process* 12(1), pp. 1–14.

Loughran, T. and B. Mcdonald (2014). “Measuring Readability in Financial Disclosures”. *Journal of Finance* 69(4), pp. 1643–1671. DOI: 10.1111/jofi.12162.

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. DOI: 10.1073/pnas.1211286109.

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. DOI: 10.1257/jep.24.2.129.

Paludi, M. A. and W. D. Bauer (1983). “Goldberg Revisited: What’s in an Author’s Name”. *Sex Roles* 9(3), pp. 387–390. DOI: 10.1007/BF00289673.

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. DOI: 10.1002/pits.1020.

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. DOI: 10.1073/pnas.1314788111.

Roberts, T.-A. and S. Nolen-Hoeksema (1989). “Sex Differences in Reactions to Evaluative Feedback”. *Sex Roles* 21(11-12), pp. 725–747. DOI: 10.1007/BF00289805.

Roberts, T.-A. and S. Nolen-Hoeksema (1994). “Gender Comparisons in Responsiveness To Others’ Evaluations in Achievement Settings”. *Psychology of Women Quarterly* 18, pp. 221–240. DOI: 10.1111/j.1471-6402.1994.tb00452.x.

Romero, J. (2013). “Where Are the Women?” *Econ Focus* 7(2), p. 12.

Sarsons, H. (2015). “Gender Differences in Recognition for Group Work”. Cambridge, Massachussets.

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.

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. DOI: 10.1073/pnas.1403334111.

Sirico, L. J. (2007). “Readability Studies: How Technocentrism Can Compromise Research and Legal Determinations”. *Quinnipiac Law Review* 26(1), pp. 147–172.

Stallard, C. K. (1974). “An Analysis of the Writing Behavior of Good Student Writers”. *Research in the Teaching of English* 8(2), pp. 206–218.

Stempel, G. H. (1981). “Readability of Six Kinds of Content in Newspapers”. *Newspaper Research Journal* 3(1), pp. 32–37.

Thörnqvist, T. (2015). “Sophistication, News and Individual Investor Trading”. Stockholm, Sweden.

Torgler, B. and M. Piatti (2013). *A Century of American Economic Review*. Palgrave Macmillan. DOI: 10.1057/9781137333056.

Van Rooyen, S., T. Delamothe and S. J. 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). DOI: 10.1136/bmj.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. DOI: 10.1136/bmj.318.7175.23.

Walsh, E. et al. (2000). “Open Peer Review: A Randomised Controlled Trial”. *British Journal of Psychiatry* 176(1), pp. 47–51. DOI: 10.1192/bjp.176.1.47.

Webb, T. J., B. O’Hara and R. P. Freckleton (2008). “Does Double-blind Review Favor Female Authors?” *Trends in Ecology and Evolution* 6(7), pp. 351–353. DOI: 10.1016/j.tree.2008.03.003.

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). DOI: 10.3389/fpsyg.2011.00178.

Weisshaar, K. (2014). “Publish and Perish?: An Assessment of Gender Disadvantage in Academia”. Stanford, California.

Whittaker, R. J. (2008). “Journal Review and Gender Equality: A Critical Comment on Budden et al.” *Trends in Ecology and Evolution* 23(9), pp. 478–479. DOI: 10.1016/j.tree.2008.06.003.

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.

Xie, Y. and K. A. Shauman (2005). *Women in Science: Career Processes and Outcomes*. Harvard University Press.

Appendix

A English as a native language

In this section, I conduct a primitive surname analysis to determine whether female economists are more (or less) likely to speak English natively.[83] To proxy for native language, I construct a binary variable equal to one if an author's last name is shared with at least 100, 1,000, *etc.* people in the U.S., according to the 2000 Census. Given historical immigration to the U.S., I supplement the analysis with an analogous indicator based on popular Scottish surnames during 1975–2015 (shared by 10 or more people); data are from the National Records of Scotland.[84]

Table A.1 displays correlations between the various surname popularity variables. Note the substantial overlap between Scottish and U.S. surnames shared by 1,000–100,000 people.

TABLE A.1: Correlation matrix of surname popularity variables

	10 (Scotland)	100 (U.S.)	$1K$ (U.S.)	$10K$ (U.S.)	$100K$ (U.S.)	$1M$ (U.S.)
10 (Scotland)	1.000					
100 (U.S.)	0.326	1.000				
$1K$ (U.S.)	0.493	0.650	1.000			
$10K$ (U.S.)	0.672	0.412	0.634	1.000		
$100K$ (U.S.)	0.585	0.218	0.335	0.529	1.000	
$1M$ (U.S.)	0.219	0.073	0.112	0.177	0.335	1.000

Notes. Sample 7,394 authors. Correlation matrix of surname popularity variables. Each variable is equal to one if a surname is shared by at least x people in the correponding population, where x is 10, 100, $1K$, *etc.* Source: U.S. Census and the National Records of Scotland.

Table A.2 displays coefficients on male$_i$ from a probit regression with the surname indicators as dependant variables. In the first column, male authors are slightly more likely to have popular U.S. and Scottish last names, although figures are statistically significant for very popular American surnames, only. Table A.2's second column includes fixed effects for the first year in which an author published in the data. Year effects are meant to control for changes in surname popularity over time—*e.g.*, due to immigration—that might be correlated with authors' gender. Indeed, their inclusion eliminates gender's impact; column two suggests men and women are equally likely to be native-English speakers.

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

[84] I use Scottish in lieu of U.K. data because only popular 1911 surnames are available from the latter. (British Census data are first publicly released 100 years after being collected.)

Table A.2: Probability authors are native English speakers, by gender

Dependant variable	(1)	(2)
10 (Scotland)	0.02	−0.08
	(0.05)	(0.05)
100 (U.S.)	0.06	−0.03
	(0.05)	(0.05)
$1K$ (U.S.)	0.02	−0.04
	(0.04)	(0.05)
$10K$ (U.S.)	0.04	−0.02
	(0.05)	(0.05)
$100K$ (U.S.)	0.12**	0.05
	(0.06)	(0.06)
$1M$ (U.S.)	0.18	0.11
	(0.13)	(0.13)
Year effects		✓

Notes. Sample 7,394 authors. Coefficients from a probit regression of male_i on a binary variable equal to one if an author's surname is shared by at least 10, 100, $1K$ *etc.* people in the listed population. Robust standard errors in parentheses. Source: U.S. Census and the National Records of Scotland. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

B Supplemental summary statistics

B.1 Average first, mean and final paper scores. Table B.3 displays authors' average readability scores for their first, mean and final papers. Grade-level scores (Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall) have been multiplied by negative one (see Section 2.1). Sample excludes authors with fewer than three publications.

Table B.3: Average first, mean and final paper scores

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Average first paper score					
Women	39.19	−13.81	−17.37	−15.19	−10.99
	(1.164)	(0.239)	(0.289)	(0.211)	(0.098)
Men	39.40	−13.76	−17.53	−15.34	−11.00
	(0.307)	(0.072)	(0.082)	(0.055)	(0.026)
Average mean score					
Women	41.16	−13.38	−16.94	−14.94	−10.91
	(0.714)	(0.145)	(0.183)	(0.134)	(0.067)
Men	39.61	−13.68	−17.41	−15.26	−11.01
	(0.186)	(0.043)	(0.048)	(0.033)	(0.016)
Average final paper score					
Women	41.86	−13.15	−16.65	−14.71	−10.90
	(1.064)	(0.219)	(0.257)	(0.185)	(0.107)
Men	39.53	−13.71	−17.41	−15.24	−11.08
	(0.325)	(0.080)	(0.090)	(0.059)	(0.026)

Notes. Sample 1,674 authors; includes only authors with three or more publications. Figures are average readability scores for authors' first, mean and last published articles. Grade-level scores have been multiplied by negative one (see Section 2.1). Standard errors in parentheses.

C Supplemental regression output

C.1 Table 4, male effects. Table C.4 shows male effects from the regressions described and presented in Table 4. Effects estimated at a female ratio of zero and observed values for other co-variates. Grade-level effects (Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall) have been multiplied by negative one (Section 2.1).

Table C.4: Table 4, male effects

	(1)	(2)	(3)	(4)	(5)	(6)
Flesch Reading Ease	39.59	39.59	39.59	39.59	39.58	40.13
	(0.037)	(0.037)	(0.038)	(0.038)	(0.039)	(0.059)
Flesch-Kincaid	−13.73	−13.72	−13.72	−13.73	−13.73	−13.48
	(0.008)	(0.008)	(0.008)	(0.009)	(0.010)	(0.013)
Gunning Fog	−17.46	−17.46	−17.46	−17.46	−17.47	−17.16
	(0.009)	(0.009)	(0.009)	(0.010)	(0.011)	(0.015)
SMOG	−15.28	−15.28	−15.28	−15.28	−15.28	−15.10
	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.011)
Dale-Chall	−11.00	−11.00	−11.00	−11.00	−11.00	−11.03
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.006)
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓
Institution effects				✓	✓	✓
Productivity effects					✓	✓
JEL effects						✓

Notes. 9,123 articles in (1)–(5); 5,216 articles in (6). Figures correspond to the male effects from regression results presented in Table 4. Effects estimated at a female ratio of zero and observed values for other co-variates. Standard errors clustered on editor in parentheses.

C.2 Table 6, male effects. Table C.5 displays total male effects—*i.e.*, the total effect for men co-authoring only with other men—from the regressions presented in Table 6. Effects estimated at a female ratio of zero and observed values for other co-variates. Grade-level effects (Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall) have been multiplied by negative one (see Section 2.1).

Table C.5: Table 6, male effects

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Male effect	39.78	−13.64	−17.37	−15.23	−11.01
	(0.151)	(0.034)	(0.038)	(0.027)	(0.013)
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓

Notes. Sample 9,166 observations (2,826 authors). Figures correspond to the male effects from regression results presented in Table 6 (first-differenced, IV estimation of Equation (1) (Arellano and Bover, 1995; Blundell and Bond, 1998)). Effects estimated at a female ratio of zero and observed values for other co-variates. Regressions weighted by $1/N_j$; standard errors adjusted for two-way clustering on editor and author (in parentheses).

C.3 Table 8 (first column), full output. Table C.6 estimates Equation (2) via OLS. The first row displays (biased) coefficients on female ratio (β_{1P}), also shown in the first column of Table 8. The second and third rows are the coefficient on score_{jW} and β_{0P}, respectively.

Table C.6: Table 8 (first column), full output

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Female ratio	1.75**	0.50**	0.53**	0.32**	0.19***
	(0.83)	(0.21)	(0.21)	(0.14)	(0.05)
Working paper score	0.84***	0.75***	0.77***	0.80***	0.84***
	(0.02)	(0.03)	(0.03)	(0.02)	(0.01)
Constant	18.20***	−2.09***	−2.09***	−1.90***	−0.59**
	(1.59)	(0.63)	(0.57)	(0.39)	(0.24)
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Year×Journal effects	✓	✓	✓	✓	✓
Productivity effects	✓	✓	✓	✓	✓

Notes. Sample 2,008 NBER working papers; 1,978 published articles. Coefficients from OLS regression of Equation (2). First row is β_{1P}, and corresponds to results presented in the first column of Table 8; second and third rows are the coefficient on score^s_{jW} and β_{0P}, respectively. Standard errors clustered on editor (in parentheses). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

C.4 Table 10, equality test statistics. Table C.7 displays χ^2 test statistics from Wald tests of β_1 (Equation (1)) equality across estimation results in Table 10.

Table C.7: Table 10, equality test statistics

	$t = 1$ vs. 2	$t = 1$ vs. 3	$t = 1$ vs. 4–5	$t = 1$ vs. ≥ 6	$t = 2$ vs. 3
Flesch Reading Ease	1.400	12.619	2.005	1.357	7.095
Flesch-Kincaid	0.084	7.547	1.098	1.882	5.431
Gunning Fog	0.303	6.530	1.183	1.840	4.690
SMOG	0.454	7.128	1.574	1.956	5.051
Dale-Chall	0.009	4.231	1.619	1.726	5.001

Notes. χ^2 test statistics from Wald tests of β_1 (Equation (1)) equality across estimation results in Table 10.

D Robustness

D.1 Table 4, alternative clustering. The following tables repeat regressions in Table 4, clustering errors instead on volume (Table D.8), issue (Table D.9) and article (Table D.10). Standard errors vary little.

Table D.8: Table 4, clustering on volume

	(1)	(2)	(3)	(4)	(5)	(6)
Flesch Reading Ease	0.91*	0.88*	0.85	0.82	0.97*	0.53
	(0.52)	(0.52)	(0.53)	(0.54)	(0.56)	(0.62)
Flesch-Kincaid	0.19	0.18	0.19	0.19	0.22*	0.23
	(0.12)	(0.12)	(0.12)	(0.12)	(0.13)	(0.14)
Gunning Fog	0.34**	0.33**	0.34**	0.34**	0.37**	0.34**
	(0.14)	(0.14)	(0.14)	(0.14)	(0.15)	(0.16)
SMOG	0.21**	0.21**	0.22**	0.21**	0.23**	0.19
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.12)
Dale-Chall	0.10*	0.10*	0.10*	0.10*	0.11**	0.10
	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓
Institution effects				✓	✓	✓
Productivity effects					✓	✓
JEL effects						✓

Notes. 9,123 articles in (1)–(5); 5,216 articles in (6). Figures estimate the effect of an article's female ratio from an OLS regression on each readability score. Standard errors clustered on journal volume in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Table D.9: Table 4, clustering on issue

	(1)	(2)	(3)	(4)	(5)	(6)
Flesch Reading Ease	0.91	0.88	0.85	0.82	0.97	0.53
	(0.57)	(0.57)	(0.58)	(0.58)	(0.59)	(0.68)
Flesch-Kincaid	0.19	0.18	0.19	0.19	0.22*	0.23
	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)	(0.15)
Gunning Fog	0.34**	0.33**	0.34**	0.34**	0.37**	0.34**
	(0.15)	(0.15)	(0.15)	(0.16)	(0.16)	(0.17)
SMOG	0.21**	0.21**	0.22**	0.21*	0.23**	0.19
	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.12)
Dale-Chall	0.10**	0.10**	0.10**	0.10*	0.11**	0.10*
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓
Institution effects				✓	✓	✓
Productivity effects					✓	✓
JEL effects						✓

Notes. 9,123 articles in (1)–(5); 5,216 articles in (6). Figures estimate the effect of an article's female ratio from an OLS regression on each readability score. Standard errors clustered on journal issue in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Table D.10: Table 4, robust standard errors

	(1)	(2)	(3)	(4)	(5)	(6)
Flesch Reading Ease	0.91	0.88	0.85	0.82	0.97*	0.53
	(0.56)	(0.56)	(0.57)	(0.57)	(0.58)	(0.67)
Flesch-Kincaid	0.19	0.18	0.19	0.19	0.22*	0.23
	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)	(0.14)
Gunning Fog	0.34**	0.33**	0.34**	0.34**	0.37**	0.34**
	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)	(0.17)
SMOG	0.21**	0.21**	0.22**	0.21**	0.23**	0.19
	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)	(0.12)
Dale-Chall	0.10**	0.10**	0.10*	0.10*	0.11**	0.10*
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓
Institution effects				✓	✓	✓
Productivity effects					✓	✓
JEL effects						✓

Notes. 9,123 articles in (1)–(5); 5,216 articles in (6). Figures estimate the effect of an article's female ratio from an OLS regression on each readability score. Robust standard errors in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

D.2 Table 4, alternative quality/productivity controls. Table D.11 repeats the regressions in Table 4 using an alternative measure of a paper's quality—the order an article appears in an issue. (For example, the lead article is assigned one, the next article two, *etc.*) It is meant to capture a paper's contribution and importance—its so-called "q-quality" (Ellison, 2002a).[85] As Table D.11 illustrates, including fixed effects for order has little impact on coefficients or their standard errors.

[85] Ellison (2002b) showed that papers published earlier in an issue spend less time in peer review. (This is supported by Table 9.) He attributes this to a "q-r trade-off": reviewers demand fewer r-quality revisions (robustness checks, clarity, *etc.*) from papers that represent important contributions (q-quality).

Table D.11: Table 4, alternative quality/productivity controls

	(1)	(2)	(3)	(4)
Flesch Reading Ease	0.83	0.95*	0.44	0.52
	(0.50)	(0.53)	(0.53)	(0.54)
Flesch-Kincaid	0.20*	0.22*	0.24*	0.23*
	(0.11)	(0.13)	(0.12)	(0.12)
Gunning Fog	0.34**	0.36**	0.34**	0.33**
	(0.13)	(0.14)	(0.14)	(0.14)
SMOG	0.21**	0.22**	0.19*	0.18*
	(0.10)	(0.10)	(0.10)	(0.11)
Dale-Chall	0.10**	0.11**	0.09*	0.09*
	(0.04)	(0.05)	(0.05)	(0.05)
Editor effects	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓
Year effects	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓
Productivity effects		✓		✓
Order effects	✓	✓	✓	✓
JEL effects			✓	✓

Notes. 9,123 articles in (1)–(2); 5,216 articles in (3)–(4). Figures estimate the effect of an article's female ratio from an OLS regression on each readability score. Standard errors clustered on editor in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

D.3 Figure 1, excluding *AER Papers & Proceedings*. Due to small samples of female authors, Figure 1 includes 561 articles from *AER Papers & Proceedings*. Figure D.1 replicates its analysis excluding these observations. In addition to dropping *JEL* codes A (General Economics, Handbooks and Teaching), B (History or Economic Thought, Methodology and Heterodox Approaches), M (Business Administration and Business Economics; Marketing; Accounting; Personnel Economics) and P (Economic Systems) (see Footnote 33), Figure D.1 excludes K (Law and Economics), N (Economic History), Q (Agricultural and Natural Resource Economics; Environmental and Ecological Economics), R (Urban, Rural, Regional, Real Estate, and Transport Economics) and Z (Other Special Topics): each has five or fewer observations that are both 100 percent female authored and assigned no more than two distinct *JEL* codes.[86] Results are consistent with those in Figure 1.

[86] Articles are assigned, on average, two distinct primary *JEL* codes. A quarter of all articles are assigned three or more primary *JEL* codes; eight percent are assigned four or more; two percent are assigned five or more (up to a maximum of eight). Observations assigned five or more *JEL* codes are disproportionately clustered in the excluded codes.

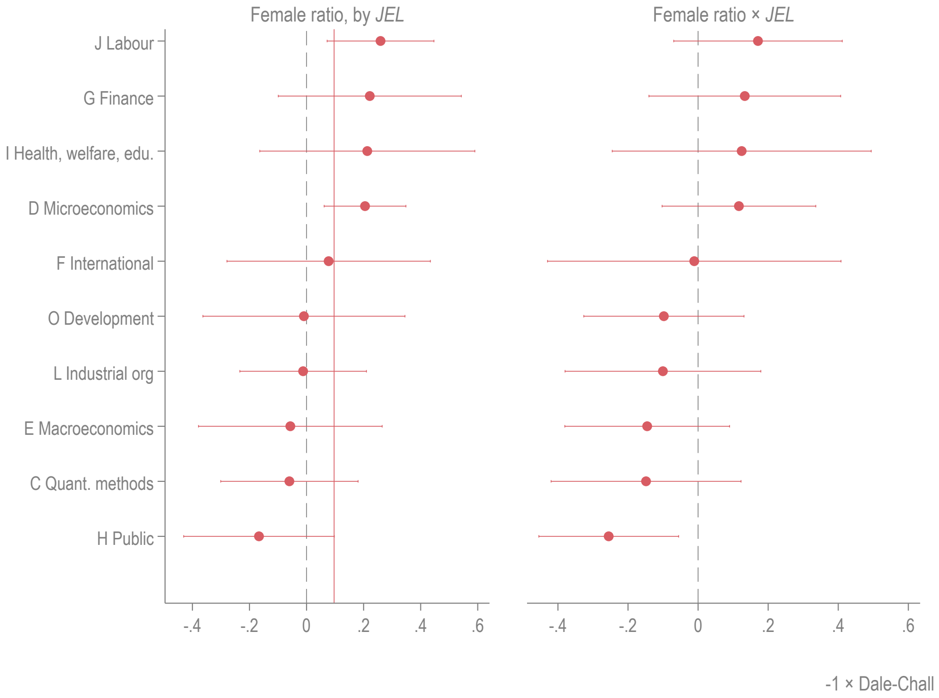

FIGURE D.1: Figure 1, excluding *AER Papers & Proceedings*

Notes. Sample 4,742 articles. Codes A, B, K, M, N, P, Q, R and Z dropped due to small sample sizes of female-authored papers (see Footnote 33 and text of Appendix D.3). Estimation strategy identical to the one in Figure 1; see Figure 1 notes for more details. Left-hand graph shows marginal effects of female ratio for each *JEL* code ($\beta_1 + \beta_3^k$); the pink vertical line is the mean effect at observed *JEL* codes (0.11, standard error 0.055). Right-hand graph displays interaction terms (β_3^k). Horizontal lines represent 90 percent confidence intervals from standard errors adjusted for clustering on editor.

D.4 Table 6, including quality/productivity controls. To remain consistent with a similar regression in Table 11, Table 6 does not include author productivity effects. Table D.12 reproduces Table 6 including these effects. Coefficients and standard errors vary little from those in Table 6.

Table D.12: Table 6, including quality/productivity controls

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Female ratio (women)	2.27**	0.35*	0.65**	0.46**	0.24**
	(1.02)	(0.20)	(0.25)	(0.19)	(0.10)
Female ratio (men)	0.81	0.13	0.18	0.11	0.12
	(1.39)	(0.26)	(0.31)	(0.22)	(0.11)
Female ratio × male	−1.45	−0.23	−0.47	−0.35	−0.12
	(1.52)	(0.31)	(0.36)	(0.26)	(0.13)
N	0.04	0.02	0.03	−0.01	0.01
	(0.22)	(0.05)	(0.06)	(0.04)	(0.02)
Lagged score	0.03*	0.04**	0.02	0.02	0.02**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)
z-test for no serial correlation					
Order 1	−20.30	−15.81	−16.75	−19.32	−21.48
Order 2	0.20	−0.57	−0.23	−0.01	−0.44
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal × Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓
Productivity effects	✓	✓	✓	✓	✓

Notes. Sample 9,166 observations (2,826 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). 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.

D.5 Table 6, equal weighting. In order to create author time series, article-level data were duplicated by article j's co-author count, N_j. Each duplicate observation was assigned a single author. Table 6 weights authors' observations by $1/N_j$—papers with fewer co-authors are weighted more heavily because they've been duplicated fewer times. Table D.13 weights all author-level observations equally. Doing so does not meaningfully affect the analysis.

TABLE D.13: Table 6, equal weighting

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Female ratio (women)	2.28**	0.33*	0.64***	0.47**	0.22**
	(0.98)	(0.17)	(0.23)	(0.18)	(0.09)
Female ratio (men)	0.73	0.14	0.19	0.13	0.12
	(1.46)	(0.27)	(0.32)	(0.23)	(0.11)
Female ratio×male	−1.55	−0.19	−0.45	−0.34	−0.10
	(1.52)	(0.32)	(0.37)	(0.26)	(0.13)
N	0.08	0.02	0.03	0.00	0.02
	(0.22)	(0.05)	(0.06)	(0.04)	(0.02)
Lagged score	0.03**	0.04***	0.03*	0.03	0.02**
	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)
z-test for no serial correlation					
Order 1	−20.46	−15.87	−16.87	−19.86	−22.54
Order 2	0.38	−0.42	−0.02	0.24	−0.25
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓

Notes. Sample 9,166 observations (2,826 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_2); female ratio (men): analogous effect for male authors ($\beta_2 + \beta_3$). z-statistics for first- and second-order autocorrelation in the first-differenced errors (Arellano and Bond, 1991) (null hypothesis no autocorrelation). Regressions unweighted; standard errors adjusted for two-way clustering on editor and author (in parentheses). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

D.6 Table 9, alternative motherhood thresholds. Table D.14 repeats the regression presented in Table 9 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 naturally vary. Remaining coefficients are unaffected.

Table D.14: Table 9, alternative mother$_j$ thresholds

	Age $<$ 3	Age $<$ 4	Age $<$ 5	Age $<$ 10	Age $<$ 18
Female ratio	5.355***	5.846***	6.030***	5.937***	5.731***
	(2.042)	(2.065)	(2.089)	(2.113)	(2.176)
Mother	−4.354*	−12.973***	−13.004***	−9.298**	−4.507
	(2.338)	(4.323)	(3.674)	(4.512)	(4.518)
Birth	2.287	10.375**	10.244**	6.636	2.057
	(3.745)	(5.179)	(4.650)	(5.335)	(5.317)
Max t	−0.132***	−0.133***	−0.132***	−0.132***	−0.132***
	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)
No. pages	0.182***	0.182***	0.182***	0.182***	0.183***
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
N	0.990**	0.974**	0.970**	0.967**	0.973**
	(0.394)	(0.393)	(0.393)	(0.394)	(0.395)
Order	0.228***	0.227***	0.229***	0.227***	0.227***
	(0.070)	(0.070)	(0.070)	(0.070)	(0.070)
Constant	37.281***	37.384***	37.389***	37.360***	37.285***
	(1.799)	(1.798)	(1.798)	(1.797)	(1.796)
Editor effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓

Notes. Sample 2,626 articles. Coefficients from OLS estimation of Equation (8) at different child age thresholds defining the binary variable mother$_j$. In column one, mother$_j$ equals one for papers authored exclusively by women with children younger than three; in column two, the threshold is children younger than four; *etc.* Column three corresponds to results presented in Table 9. Robust standard errors in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.