Author gender affects the rating of academic articles: Evidence from an incentivized, deception-free laboratory experiment.
Author gender affects the rating of academic articles: Evidence from an incentivized, deception-free laboratory experiment.

MICHAL KRAWCZYK
Faculty of Economic Sciences, University of Warsaw
e-mail: mkrawczyk@wne.uw.edu.pl

MAGDALENA SMYK
Faculty of Economic Sciences, University of Warsaw
e-mail: msmyk@wne.uw.edu.pl

Abstract
In this study we sought to verify the hypothesis that a researcher’s gender affects evaluation of his or her work, especially in fields in which women are a small minority. To this end we asked a sample of economics majors to rate papers written by mixed-gender couples, indicating that they were (co-) authored by a “female economist”, “male economist”, “young female economist” or “young male economist”. While the age factor played no role, female authors received lower ratings. This effect was independent of the subject's gender.

Keywords:
women in science, laboratory experiment, gender bias

JEL:
C91, J16
1 Introduction

Does gender still matter when we picture a successful scientist? A hundred years ago portraying a researcher as a male was quite natural, given that there were so few women in science. During the last few decades, the number of female researchers has increased dramatically, but gender imbalance remains outstanding at the higher levels of academia. Even though women now receive as many PhD diplomas as men, they hold only 19.4% and 7.9% of full-time professorships in social sciences and engineering and technology, respectively, in the European Union. Which of these two speeds of transition shapes the predominant view of women in academia?

More than 50 years ago, Mead and Metraux (1957) asked US high school students to write an essay about a scientist, and the vast majority of them described a man. A study with kindergarten children more than two decades later showed the same result: only 28 girls among some 4000 children drew a female (Chambers 1983). Follow-up research based on DAST (Draw-A-Scientist Test) methodology confirmed that most students at each stage of education had and still have a stereotyped perception of a scientist’s gender (Fort and Varney 1989, Rosenthal 1993, Steinke et al. 2007). Interestingly, younger children are more likely to draw a woman in the test (Barman 1996, Liu et al. 2010). For more than half a century of research based on DAST, results have been stable - the stereotype was (and still is) strongly noticeable irrespective of the changes that have occurred in the scientific world.

The 'male-scientist' stereotype is not only an American domain - in studies in England (Maoldomhnaigh and Hunt 1988), Korea (Song and Kim 1999), China (Liu et al. 2010) and Turkey (Turkmen 2008), most children drew men. This view is also unrelated to being a child or a student - adults also consider it much more likely that a scientist will be a man rather than a woman (Losh 2010). Once a fact of life, it is now an obsolete stereotype that is possibly harmful to women considering or pursuing a scientific career and looking for role models. As Nosek et al. (2009) claim, the 'male-scientist' image creates a barrier for women to become successful scholars. Women may believe that they are less talented (especially in math), and they will consequently be less willing to study ‘male-type’ subjects. Therefore, they will less frequently choose a career in science compared with men. Importantly, Nosek et al. (2009) point out that the ‘male-scientist’ stereotype might be easily transformed to ‘male-better scientist’ without any factual basis. Unconscious belief in correlation between gender and qualification may influence evaluation of scientific achievements and impede development of female researchers.

This is particularly important given that scientists are almost constantly being evaluated. With submissions to journals and conferences, applications to schools, and job interviews, even a trace of prejudice in the competitive world of academia might have a lasting impact on an individual’s future career.

The literature provides some evidence that gender bias in academia exists. Wenners and Wold (2001) found that female postdocs had to publish several more articles in top journals (and over a dozen in less known ones) to be judged as productive as similar male candidates during the recruitment process. Women also receive fewer grants and smaller funds for research compared with male scientists (Bornmann et al. 2007). Studies on journal publication history show that men are favored as papers’ authors in several disciplines such as ecology (Tregenza 2002), medicine (Link 1998) and biomedicine (Wenners and Wold 2001). Men also win many more awards in STEM disciplines than would be predicted by the proportion of nominees and their position in academia (Lincoln et al. 2012). On the other hand, Canizares (2009) pointed out that in the last 20 years at many American research-intensive universities, women have been more likely to be invited to job interviews and to receive a position than men. Likewise, Budden et al. (2008) indicated that even if introducing double-blind peer review (i.e., the identities of both the author and the reviewer are hidden) increased representation of female authors in biological sciences, reasons other than gender bias could explain the phenomenon (Engqvist and Frommen 2008). See Ceci and Williams (2011) for a critical review of this research.

Extracting differences between men and women from existing data is not the only way to study gender bias in academia. Experimental designs allow creating circumstances in which ‘evaluated scientists’ differ only by gender. In a randomized experiment Moss-Racusin et al. (2012) found that male students were more likely to be hired as lab manager at research universities than their female colleagues. Mediation analysis showed this was due to female candidates being judged less competent, although by design the applications showed equivalent experience and skills. Borsuk et al. (2009) manipulated a published paper
to give it four different author designations (female, male, initials, and no name at all). It was reviewed by potential referees at four stages of academic training: undergraduates, graduate students, postdoctoral researchers and faculty researchers. A request to judge if the paper was suitable for publication was sent by e-mail, and subjects were asked to review the abstract online, with the exception of biology undergraduates who completed the review in class. No gender bias was observed; rejection rates and quality ratings were not affected by the experimental manipulation. However, female postdoctoral referees were the most critical and were more likely to reject the paper compared with other groups.

In a similar study, Knobloch-Westerwick et al. (2013) conducted an experiment in which communication scholars evaluated abstracts taken from the International Communication Association’s 2010 conference. The topics were either gender-typed or gender neutral, and some abstracts were ostensibly signed with female or male names. The study was performed online, and subjects also completed a questionnaire on their opinions about gender roles. Among six targeted abstracts, ‘male-authored’ texts were rated higher on average, especially when the topic was male-typed. The gender of the participant was insignificant.

Our experiment expanded on these designs, with modifications partly dictated by the experimental paradigm of our discipline. To begin with, we did not mislead our subjects—they were really evaluating papers written by professional male and female economists. This is of importance, if only because deception may raise suspicion (Davis 1993, Ortmann and Hertwig 2002). Second, we gave our subjects the entire paper, rather than just the abstract or a very short text, and ample time to study it. Third, we provided conditions (a laboratory without Internet connection) that practically eliminated the possibility of uncontrolled consulting of external sources. Fourth, we asked several questions about various elements of the paper. Fifth, we added the age dimension on the suspicion that young female researchers may be evaluated particularly harshly. Finally, to the best of our knowledge, this was the first experimental study on reviewing scientific articles in the field of economics.

2 Experimental design and procedures

Before the experiment, we selected five pairs of economic papers (see Appendix B). Paired papers were written by the same mixed-gender couple of authors and came from top schools (such as Harvard, Columbia and MIT). One paper within each pair had already been published in a highly rated economic journal (The Quarterly Journal of Economics, Journal of Economic Growth, or Journal of Development Economics), while the other was a working paper yet to be published. Both groups showed similarly high scientific quality.

This similarity was tested in a pilot study. Four students (three third-year economics majors and one PhD candidate in sociology) read all 10 papers and answered questions concerning their intelligibility and academic excellence (rated on scale from 1 to 10) as well as the time needed to read the text. There were no significant differences in the ratings of the two types of papers (published and unpublished).

In the experiment, subjects were asked to read one paper from our pool and evaluate its scientific quality and the competence of the author. We asked them for a general opinion on the author’s competence, language quality of the paper, study methodology, comprehensiveness of literature review, significance of obtained results, general quality of the text and its intelligibility (see Instruction—Appendix A). In each case the subject had to make a rating on a 0-7 scale. Finally, we asked whether the subject thought the paper was published or not (knowing the base rate of 50%).

We applied four treatments based on two conditions (two by two design)—one related to the gender of the author and the second to his or her age. Indeed, papers were described as written by “a female economist”, “a male economist”, “a young female economist” or “a young male economist”. For the sake of brevity, we shall refer to the papers as male-authored or female-authored henceforth (depending on the information received by subjects). It should be noted that in the Polish language (in which the study was conducted) distinguishing the gender of an economist (“ekonomista” vs. “ekonomistka”) occurs readily. All four treatments were applied to each paper (with a slight imbalance resulting from different numbers of subjects showing up for each session).

1We did not, however, give our subjects any actual names, to exclude the possible effect of nationality/ethnicity. The papers were provided in a form that prevented differentiating which was published and which was not (e.g., by looking at editing). The subjects were also explicitly forbidden to consult Internet sources, and we did not record any such attempt.
In the invitation to the experiment, we stressed that we were looking for participants familiar with economics and proficient in English. Subjects were given 45 minutes to read the text and answer the questions, for which they received 80 zlotys (about 20 EUR, 25 USD). Subjects who correctly guessed whether the paper was published in one of the leading economic journals or not published at all, received 10 zlotys more. Only two papers were used in each session—the published and the unpublished one written by the same authors. In one session both papers were matched only with one author’s gender to reduce the probability that subjects would discover differences between questionnaires. The same two papers would be used in another session with the authors’ genders swapped.

A total of 193 subjects took part in 10 sessions at the Laboratory of Experimental Economics at the University of Warsaw (Poland) in November and December 2013. Over 60% of the subjects were women. Nearly two-thirds of the subjects were students, and 85% of them studied economics. There were about 30-40 students from each year of study except the first year—there were only nine freshmen. The subjects’ average age was 22 years.

2.1 Predictions

The design allowed testing of the following hypotheses:

**Hypothesis 1** Female authors will receive lower ratings than males, and their papers will less often be judged as published.

**Hypothesis 2** Likewise, young economists will indirectly be judged as inferior with regard to those of unspecified age.

We also speculated that answers to specific questions may be susceptible to gender bias to different degrees. To the extent that subjects may wish to hide their prejudice unless incentivized to do otherwise, the gender effect may have been greater for the question of whether the paper was published and less for the remaining questions. Furthermore, for some dimensions that are associated with the female gender, notably language quality or comprehensiveness of literature review, the gap may have actually been reversed (Spear 1984, Eccles and Blumenfeld 1985, Sadker and Sadker 2010).

---

2The experiment was followed by another, unrelated design, in which subjects could lose some of their money. Mean earnings were some 65 zlotys, which is more than would normally be expected for a session of about 2 hours.
3 Results

To set the stage for our analysis, we must first note that judging whether a paper was published or not turned out to be very difficult for our subjects—the success rate was not significantly higher than the pure chance baseline of 50%, even for economics majors in the sample. 

Overall, 59.8% of male-authored papers were judged as published, while the figure stood at 45.8% for females-authored ones. The conservative approach to testing for significance of this treatment effect is to consider each paper as one observation, checking whether it was more often judged as published when it was “written” by a male or a female.

Figure 1: Less subjects believed that female-authored papers had been published

![Figure 1](image)

Figure 1 shows, for each of the 10 papers, the fraction of subjects who thought it had been published, depending on whether they were told that the author was female or male. Quite tellingly, for every single paper this fraction was higher in the latter case. This is obviously highly significant (\(p < 0.002\) in a two-sided test).

Table 1: Differences in ratings of ‘male-authored’ and ‘female-authored’ papers were insignificant

<table>
<thead>
<tr>
<th></th>
<th>male author average</th>
<th>female author average</th>
<th>Two-sided MWW test (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>author’s competence</td>
<td>5.15</td>
<td>4.99</td>
<td>0.17</td>
</tr>
<tr>
<td>language quality</td>
<td>4.98</td>
<td>4.92</td>
<td>0.79</td>
</tr>
<tr>
<td>proper methodology used</td>
<td>4.83</td>
<td>4.77</td>
<td>0.62</td>
</tr>
<tr>
<td>comprehensiveness of literature review</td>
<td>4.76</td>
<td>4.66</td>
<td>0.73</td>
</tr>
<tr>
<td>scientific significance of results</td>
<td>4.08</td>
<td>4.17</td>
<td>0.76</td>
</tr>
<tr>
<td>overall scientific quality</td>
<td>4.70</td>
<td>4.56</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Responses on 0-7 scale

The (unincentivized) ratings of specific aspects of the text and the author him- or herself tend to be higher for male-authored papers, but differences were not significant, see Table 1. Thus Hypothesis 1 and the conjecture that incentives intensify the effect are confirmed.

Table 2: Ratings were highly correlated with the judgment whether paper was published

<table>
<thead>
<tr>
<th></th>
<th>author’s competence</th>
<th>language quality</th>
<th>methodology</th>
<th>literature review</th>
<th>scientific significance</th>
<th>scientific quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>judged published</td>
<td>0.476***</td>
<td>0.3691***</td>
<td>0.4152***</td>
<td>0.3968***</td>
<td>0.4121***</td>
<td>0.4975***</td>
</tr>
</tbody>
</table>

*** \(p < 0.01\)

Two interpretations of the differentiated treatment effect seem natural. First, it could be that evaluations were unaffected by gender, but subjects believed that (perhaps due to gender discrimination),

---

3This did not seem to follow, however, from subjects’ inability to understand the papers—the average intelligibility rating was 4.1 (0-7 scale).
a male-written paper was more likely to be published, regardless of its qualities. Second, as mentioned before, it could be that incentives applied only to the question of publishability encouraged subjects to fully reveal their bias there. Very high correlations between the rating of various qualities and the judgment of whether the paper was published (see Table 2) suggest that the latter may be closer to the truth. If so, incentives could be an important design feature in experiments looking for gender bias in the evaluation of academic work.

Figure 2: Gender of the subject had little impact on rating

We also wondered if the effect of author’s gender was stronger for male or female subjects. As we can see in Figure 2, guessing in favor of a male economist was independent of the subject’s gender. We thus found no evidence of either ‘women against women’ or sexism/same-gender preference effect (Paludi and Bauer 1983, Lloyd 1990).

An analogous analysis was performed for the second manipulation—that of authors’ implied age. Yet Hypothesis 2 was clearly rejected—no difference was found for any group of readers. The interaction between female and young age factors was also insignificant.

Table 3: Paper judged as published: probit regressions

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>female author</td>
<td>-0.353*</td>
<td>-0.350*</td>
<td>-0.370*</td>
<td>-0.462*</td>
</tr>
<tr>
<td></td>
<td>(0.182)</td>
<td>(0.184)</td>
<td>(0.212)</td>
<td>(0.261)</td>
</tr>
<tr>
<td>young author</td>
<td>0.146</td>
<td>0.147</td>
<td>0.133</td>
<td>0.143</td>
</tr>
<tr>
<td></td>
<td>(0.182)</td>
<td>(0.184)</td>
<td>(0.199)</td>
<td>(0.200)</td>
</tr>
<tr>
<td>female subject</td>
<td></td>
<td></td>
<td>-0.561***</td>
<td>-0.428</td>
</tr>
<tr>
<td></td>
<td>(0.222)</td>
<td></td>
<td>(0.305)</td>
<td></td>
</tr>
<tr>
<td>year of study</td>
<td></td>
<td>0.00341</td>
<td></td>
<td>-0.000533</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0014)</td>
<td></td>
<td>(0.0018)</td>
</tr>
<tr>
<td>not a student</td>
<td></td>
<td></td>
<td>-0.290</td>
<td>-0.315</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.418)</td>
<td>(0.420)</td>
</tr>
<tr>
<td>female author*male subject</td>
<td></td>
<td></td>
<td></td>
<td>0.255</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.424)</td>
</tr>
<tr>
<td>paper effects included</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>observations</td>
<td>193</td>
<td>193</td>
<td>173</td>
<td>173</td>
</tr>
<tr>
<td>pseudo $R^2$</td>
<td>0.017</td>
<td>0.039</td>
<td>0.0759</td>
<td>0.0774</td>
</tr>
</tbody>
</table>

Paper effects was insignificant at 1%.

Subjects who took part in the first session are not included in third specification, because due to a technical error their demographic data is missing.

Standard errors in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

For a comprehensive assessment of factors affecting whether a paper was judged as published or not, we performed probit regressions using various specifications, see Table 3. Not surprisingly, our findings
concerning treatment effects were confirmed—female gender of the author was statistically significant, while the youth factor was not. Female subjects turned out to be less likely to recognize a paper as having been published (which corresponds to one of results reported by (Borsuk et al. 2009), but again, there was no interaction between the author’s and the subject’s gender. We also checked that the gender effect did not depend on the year of study (a proxy for competence in evaluation of scientific quality). Paper fixed effects, whenever included, were jointly insignificant.

4 Discussion and conclusions

The main finding of our study is that the gender of the author, but seemingly not his or her age, matters when it comes to the evaluation of a paper, at least in the field of economics. The fraction of guesses that a male-authored paper had been published was 30.6% (14 pct. points) higher than for female-authored ones. As with any laboratory experiment, the question that arises is to what extent this tendency carries over to contexts of the greatest practical relevance, such as peer review and career choices.

Clearly, one would wish to conduct such a study with active researchers rather than students, but it is prohibitively costly to arrange meaningful incentives (and prevent the use of the Internet) with such a sample. It must be emphasized, however, that our subjects were not simply a random collection of college students. The University of Warsaw is the best in the nation as is its economics curriculum. Additionally, as mentioned before, it was stressed in the invitation that the task would involve reading a scientific paper in English (and sessions were scheduled for late afternoons!), likely leading to substantial self-selection: the subjects not only had to possess the required skills, but also had to be strongly interested in social sciences. It is thus likely that quite a few of them will indeed pursue an academic career in the near future (and the chances are that their views of the role of women in academia will not change dramatically until then). Additionally, it seems plausible that students’ perceptions of the issue are being shaped (perhaps unconsciously) by those of their academic teachers—in this indirect sense, the observed results may be informative of the beliefs prevailing in academia as well. It is worth mentioning in this context that the University of Warsaw is, in terms of the participation of women in science, representative of Europe - about 50 to 60% of PhD candidates and only about 20% of full-time professors are women.

Note also that our subjects were approaching the age at which important early career decisions are made and when inexperienced researchers may easily be discouraged. From this perspective, young people’s views may be of greater importance than those of older adults. In particular, our finding that the author’s gender effect was just as large in female as in male subjects, suggests that women fail to see successful female economists that could serve as role models. If they believe that women are simply not as good as men in science or that women in academia encounter a lot of obstacles, they will likely choose a different field to work in.

Perceptions of gender roles among the young could be an indicator of the changes to come. In view of our results, the outlook for female economists is not particularly bright.
References


Canizares, C. R.: 2009, Gender differences at critical transitions in the careers of science, engineering, and mathematics faculty, *In Memoriam* p. 35.


Lloyd, M. E.: 1990, Gender factors in reviewer recommendations for manuscript publication, *Journal of applied behavior analysis* 23(4), 539–543.


Welcome to the Laboratory of Experimental Economics! This session will consist of two mutually unrelated parts. In the first part you will earn money. In the second one you may lose either part or all of the money you earn in the first part. It is obligatory to take part in both. Immediately after the session we will pay out your earnings in cash.

Please turn off your mobile phone and any other electronic devices you have. Do not talk or communicate with other participants during the session. Those who disobey these rules will be excluded from experiment without remuneration. If you have a question or doubt, please raise your hand. The experimenter will come to you.

Both experiments are conducted for scientific purposes. Your answers will be treated as confidential. Results will be presented only in aggregated form.

PART I

In the first part your task will be to evaluate competence and scientific prospects of a female economist/a male economist/a young female economist/a young male economist. Please, read one of her / his papers. When you are done, please answer a few question about the text and the author. In particular, we would like you to evaluate scientific quality of the text and guess if it was:

- a) published (perhaps in an abbreviated version) in one of the top economic journals, or
- b) not published at all.

You should know that we used the same number of published and unpublished papers in this experiment.

We expect that you read the text and complete the survey in 45 minutes. For this effort we will pay you 80 zlotys. Additionally, if you correctly indicate a) or b), you will get 10 zlotys more. We will let you know if your answer was correct or not at the very end of the session (we need time to verify it).

[At this point subjects were asked to read the paper]
Answer the following questions on a scale from 0 (very low) to 7 (very high):

1. How would you evaluate author's competence?

2. How would you evaluate the text in terms of:
   a) language quality?
   b) methodology?
   c) comprehensiveness of literature review?
   d) significance of results?

3. How would you evaluate overall (?) scientific quality of the text?

4. How would you evaluate intelligibility of the text?

5. To the best of your judgment, was this text:
   a) published (perhaps in an abbreviated version) in one of the top economic journals,
   or
   b) not published at all?

   If you indicate correctly, you will earn additional 10 zlotys. When you are done, wait for further instructions. Remember, you cannot communicate with other participants during the session.
## Appendix B: List of articles used in the experiment

<table>
<thead>
<tr>
<th>Female author</th>
<th>Male author</th>
<th>Published paper</th>
<th>Unpublished paper</th>
</tr>
</thead>
</table>