CSEF

Centre for Studies in Economics and Finance

WORKING PAPER NO. 418

Research Quality and Gender Gap in Research Assessment

Tullio Jappelli, Carmela Anna Nappi and Roberto Torrini

October 2015

University of Naples Federico II

University of Salerno

Bocconi University, Milan

CSEF - Centre for Studies in Economics and Finance
DEPARTMENT OF ECONOMICS – UNIVERSITY OF NAPLES
80126 NAPLES - ITALY
Tel. and fax +39 081 675372 – e-mail: csef@unina.it

CSEF
Centre for Studies in Economics and Finance

WORKING PAPER NO. 418

Research Quality and Gender Gap in Research Assessment

Tullio Jappelli[*], Carmela Anna Nappi[**] and Roberto Torrini[***]

Abstract
The literature on the gender gap in science reveals differences in wages, productivity, access to funding and impact on the scientific community that disadvantage women. This paper contributes to work on the gender gap in science by investigating issues such as the presence of differences in research quality between genders, the effect of family responsibilities on research quality, differences in collaborations and international co-authorships, the effect of evaluation methodology, i.e. whether bibliometric evaluation disadvantages women, and the presence of discrimination defined by referees' gender. We use the data from the National Research Assessment (VQR 2004-2010) conducted by the Italian Agency for the Evaluation of Universities and Research Institutes. These rich data allow us to control for individual variables, research output characteristics and university and scientific sector fixed effects. We find that gender differences in research quality are reduced if we control for researchers' observable characteristics, evaluation method, and referees. In particular, we find that maternity and the intensity of research collaborations and international co-authorships play no role in explaining research quality differences. Further analysis of a random sample of papers evaluated using bibliometric indicators and peer review reveals that bibliometric evaluation does not penalize women with respect to men.

Keywords: Gender Gap, Research Evaluation.

[*] Università di Napoli Federico II and CSEF. Postal address: Department of Economics and Statistics, Via Cinthia 26, 80126, Napoli, Italy; email: tullio.jappelli@unina.it

[**] ANVUR, Via Ippolito Nievo 35, 00153, Rome, Italy; email: carmelaanna.nappi@anvur.it.

[***] ANVUR, Via Ippolito Nievo 35, 00153, Rome, Italy; email: roberto.torrini@anvur.it.

Table of contents

1. Introduction

Gender gaps in the labor market are a key policy issue in European countries. Despite EU adoption in 2000 of workplace legislation which prohibits discrimination on the basis of racial or ethnic origin, religion or belief, disability, age, or sexual orientation, labor economists observe persistent gaps in labor market participation and wages.

The gender gap in science is of particular interest, given the role of science in promoting technological developments and economic growth. This gap can manifest itself in various ways. The literature provides evidence of a productivity gap (Mauleón and Bordons, 2006; West et al., 2013; Larivière et al., 2013), and gaps in access to finance (Ley and Hamilton, 2008), impact on the scientific community (Lariviere et al., 2011; Larivière et al., 2013), and promotions (Bagues et al., 2014) and concludes that differences along all these dimensions disadvantage women. There are several reasons ranging from time spent in child-rearing, discrimination, and stereotypes about innate talent which account for the gender bias in science and other segments of the labor market (Leslie et al., 2005).

This paper contributes to the literature on the gender gap in science by addressing several important issues. The first step in our analysis measures the gender gap using a large dataset of around 180,000 research papers published in 2004-10. Evaluation of their research quality was conducted by the Italian Agency for the Evaluation of Universities and Research Institutes (ANVUR) with the help of 450 researchers distributed across 14 panels, and about 15,000 referees. Research quality is measured by two indicators: (1) each paper's quality score (or other research output such as patents), and (2) the probability that the paper was evaluated as “excellent”, receiving a top score of 1. The richness of our dataset allows us to control for researchers' characteristics (age, gender, university affiliation, rank, and scientific subject area) and research output (type of research output, number of authors, international co-authorship, and language). Work on the gender gap focuses on wage differentials and explains the gap in terms of differences in productivity and/or discrimination. We have the opportunity to explore the gap in a context where salaries are identical for both genders, given the same age and rank. By controlling for age and rank (assistant, associate, full professor) we control automatically also for professors’ salaries since in Italy, academic contracts are based on public sector contracts, applied to all universities and research centers.

The second step in our analysis digs deeper into the various reasons that might explain the gender gap, focusing on external constraints (family duties) and discrimination, and considering the gender of the referees, and the evaluation method. To take account of family duties, we check whether the gap is wider among woman who have taken maternity leave. Using the entire sample of peer reviewed papers we control also for the characteristics of referees (in particular, whether papers were refereed by males or females), and test whether the referee's gender affects the outcome of the peer review process. VQR's research evaluation is based in part on bibliometric indicators and in part on peer review. This allows us to compare bibliometric evaluation and peer review for a random sample of papers evaluated using both methods. The objective is to check whether either of the two methods disadvantage women, for instance because citation metrics are gender biased (Ward et al., 1992; Davenport and Snyder, 1995, Larivière et al., 2013; HCEFE, 2011).

In our baseline estimates, with no further controls, we find that men's scores are significantly higher (by 5 percentage points) than women's score, and that males are about 7 percentage points more likely to receive top evaluations. However, when the quality score is the dependent variable, gender inequality falls sharply (to 1%) if we control for observable variables such as age, university rank, subject area, and university dummies, while the gap persists in the regressions for the probability of receiving the highest score.

Maternity leave does not contribute to explaining research quality differences; i.e., research quality for women who have experienced maternity leave is no different from the research quality for women (or men) who have not taken parental leave. We find strong evidence that research collaborations – measured by number of authors - and international co-authorships tend to be associated with higher quality research but neither characteristic affects the gender gap in research. Further analysis based on the sample of peer reviewed papers finds very little evidence of a "same sex preference", and analysis of a random sample of papers evaluated by bibliometric analysis and by peer review suggests that bibliometric evaluation is not significantly associated with authors' gender, and hence does not disadvantage women. Finally, focusing on single-authored papers, we rule out that gender discrimination occurs at the stage in which papers are selected by departments before the evaluation.

Overall, if we control for authors' observable characteristics, evaluation methods, and referees, some evidence emerges of a persistent (albeit small) gender gap in research, and

particularly for the smaller fraction of papers that received the highest evaluations. A possible explanation for the gap for top papers might be due to a "selection effect" rather than an "evaluation effect": women tend to choose slightly less risky projects and research strategies which results in a lower fraction of top papers.

The rest of the paper is organized as follows. Section 2 describes the main features of the Italian research evaluation, section 3 presents the data, and section 4 presents the baseline empirical estimates. The effects of maternity leave, referees' characteristics and evaluation methods are discussed in sections 5, 6 and 7, respectively. Section 8 contains an analysis for single-authored papers. Section 9 summarizes our main findings.

2. Italian research evaluation

The Italian research evaluation exercise (or VQR) was carried out between end 2011 and July 2013 by the National Agency for the Evaluation of University and Research (ANVUR).[1] It involved around 180,000 articles, books, patents, and other research output (in what follows, we use the umbrella term research papers to refer to all these types) published or produced between 2004 and 2010, and submitted by Italian universities and research bodies. The purpose of the evaluation was to rank research institutions and departments within each research area based on the quality of their research.[2]

The evaluation was performed by 14 panels - for each broad research area. Each panel included an average of 32 researchers. Papers were evaluated based on bibliometric indicators (a combination of the journal impact factor and number of citations received by each paper), or "informed" peer review by two external referees.[3] Peer review was "informed" because the papers involved had been published between 2004 and 2010 rather than being anonymous manuscripts submitted for publication and evaluated by anonymous referees. Therefore, the reviewers were aware of the author's name, gender, and affiliation. Typically, peer review

[1] ANVUR was established by a Presidential Decree (PD) published in February 2010. ANVUR's mission is to evaluate the research and study programs of Italian universities.

[2] The 14 research areas are: (1) Mathematics and Computer Sciences, (2) Physics, (3) Chemistry, (4) Earth Sciences, (5) Biology, (6) Medicine, (7) Agricultural and Veterinary Sciences, (8) Civil Engineering and Architecture, (9) Industrial and Information Engineering, (10) Ancient History, Philology, Literature and Art History, (11) History, Philosophy, Pedagogy and Psychology, (12) Law, (13) Economics, Business and Statistics, (14) Political and Social Sciences.

[3] Journal articles dealing with multidisciplinary or innovative issues on the border of different panels were evaluated by peer review.

evaluation was carried out by two external and independent reviewers, chosen by the panel members taking account of conflicts of interest. In the case of diverging assessments, a third peer review or a consensus group from among the panel could be operationalized in order to agree a synthetic and final score.[4]

The mix of informed peer review and bibliometric evaluation varied according to the research area, with an overall constraint (defined by the VQR Call), that at least 50% of the papers must be evaluated by peer review.[5] In practice, bibliometric evaluation was used quite extensively in scientific areas such as chemistry, physics, biology, medicine, where most papers are published in journals, and where most journals are indexed in ISI Thompson Reuters or Elsevier databases.[6] Peer review was almost exclusively used in areas such as Arts and Humanities, History, Law, and Social Sciences where many publications are in the form of monographs and book chapters, and bibliometric databases are incomplete or missing. In the cases of economics, business, and statistics evaluation by peer review and bibliometric indicators was split fairly evenly.

Moreover, a random 10% sample of the papers evaluated by bibliometric indicators (in hard sciences and economics) was also evaluated by peer review.[7] This implies that for a subset of about 7,000 papers we have results for evaluation by both methods which allows us to explore the potential effects of the evaluation method on the gender gap.

Researchers affiliated to an Italian university submitted their three best papers for evaluation, and researchers affiliated to public research centers submitted six papers. Each was given a score ranging from 0 to 1. Papers classified "in the top 20% of the quality ranking shared by the international scientific community" received a score equal to 1, papers in the 60%-80% range scored 0.8, papers in the 50%-60% range scored 0.5, and papers below the median received a score of zero. Each department's total score was computed as the average score of all papers submitted by the department.[8]

[4] For further details regarding VQR 2004-2010 exercise and assessment methodologies please see Ancaiani et al. (2015).

[5] 53% of papers were evaluated by peer review.

[6] In these areas, papers sent for peer review were papers published in journals not indexed by the main databases, and papers for which bibliometric indicators were not reliable and/or were consistent (e.g. papers published in 2010 for which available citations at the time of the research evaluation referred to only one year). We replicated our analysis excluding these areas and focusing only on areas where all papers were evaluated by peer review, and found the same results as reported in this paper.

[7] In practice, hard sciences correspond to Areas 1-9 (see fn. 2).

[8] In the VQR, each department's average score also reflects penalties for missing papers (-0.5), non admissible papers (-1) (e.g.published before 2004 or after 2010), and cases of fraud or plagiarism (-2). Here, we focus on

3. The data

Our sample includes data on almost 180,000 papers published in 2004-10 and submitted in early 2012 by universities and research centers to the VQR.[9] For each paper, we merge publication data (publisher, type of publication, number of authors, international co-authorship, language of publication, and evaluation method) with data on researchers' characteristics (age, gender, affiliation, rank, scientific area). For papers evaluated by peer review, we have data on the gender of the two referees. The dataset also includes the outcome of the evaluation in terms of the final score, a number ranging from 0 to 1.

Table 1 reports means and standard deviations of the variables used in the estimation by gender, and for the total sample. Males submitted 118,949 papers, or about two thirds of the sample, with the remaining third authored by women.[10] The sample includes 13% of papers submitted by relatively young researchers (less than 40 years old), 55% submitted by researchers aged between 40 and 55 years, and 32% by older researchers (aged over 55 years). While women are well represented in the younger age group, the fraction of papers authored by women declines with age (in the oldest group the fraction of papers authored by males is 9 percentage points higher than for females). This pattern reflects a strong cohort effect on access to an academic profession, with the access to academic positions by women increasing quite significantly over time.

Most papers submitted to the VQR are published in journals (74%) but there are also several book chapters (11%), monographs (8%), and other research outputs (7%). Overall, women submitted a lower fraction of journal articles and more book chapters compared to men. In our sample, male authored papers have a higher probability of international co-authorship (23% vs. 19% for females). On average, 26% of the papers submitted by women are written in Italian, while for males the fraction is 21%. The proportion of single-authored papers is 25% for males against 31% for females. A higher proportion of papers submitted by males was evaluated by bibliometric indicators, reflecting the higher proportion of journal

---

those papers that were actually evaluated, and therefore drop observations with negative scores (missing and non admissible papers).

[9] The VQR requires universities to submit 3 papers for each researcher. Missing submissions are assigned a negative score (-0.5); we exclude observations with negative scores. Note that the proportion of papers with non-negative scores is slightly lower for women (94.8%) than for men (95.1%).

[10] In the case of papers with more than one author, the gender is based on the researcher submitting the paper. If a paper has more than one author, the gender of the submitting author might not be same as the gender of the other co-authors. We checked the robustness of the results controlling for the number of authors of each paper, and limiting the sample to single authored papers.

articles submitted by men. In relation to the distribution of positions by gender, women are far less well represented among full professors, who tend to be older than associate and assistant professors.

Table 1 also reports the distribution by gender of our two quality indicators. The average score of papers submitted by men is 0.66, against 0.63 for women, resulting in a gender gap (ratio between the two scores) of 4.8%. Table 2 reports the distribution of papers in the four merit classes defined by the VQR: excellent, good, acceptable, limited. The fraction of papers in the top class is significantly higher for men (39% vs. 32% for women). Correspondingly, the fraction of papers in the three lower merit classes is higher for women. This implies that in our data gender inequality in the research gap is almost entirely dependent on differences in the upper part of the quality score distribution.

Table 3 reports the average quality score and the percentage of excellent papers, for all research areas, and for the total sample by gender, and the statistical significance of the gender difference. At the aggregate level the difference in the average quality score is 0.03 and is statistically significant at the 1% level. The pattern is similar for all research areas except Medical Science, Civil Engineering, and Psychology. The areas with the largest gender gap are Mathematics and Computer Sciences, Social Sciences, Humanities, Law and Biology. It is interesting that in Civil Engineering and Industrial and Information Engineering, which have the lowest presence of women among all the areas, women score higher than men.

In the total sample the fraction of excellent papers is 7 percentage points higher for men and the difference is statistically significant at 1%. There is a similar gap in all research areas except Engineering (Civil and Industrial) and Veterinary Sciences, where the sign of the difference is inverted in favor of women, and in the latter case, statistically significant at the 1% level. According to this indicator, Mathematics and Computer Sciences, Biology and Humanities are the research areas with the highest gender gap.

4. The determinants of research quality

In this section we check whether the unconditional difference in performance between men and women changes if we control for the observables characteristics of papers and authors. In the first specification of Table 4 we report a linear regression where the dependent variable is the quality score of each paper (ranging from 0 to 1) and the independent variables

are a dummy for gender and two age dummies. The regression also includes the full set of 367 dummies for scientific sector, and 129 dummies for the university or research institution to which authors are affiliated. We find that being female reduces the quality score assigned to the paper by 4 percentage points, and that the coefficient is statistically significant at the 1% level (column 1). Also, younger researchers have higher scores than the excluded age class (over 55 years).

In the second regression in Table 4, we control also for the observable characteristics of the research output using dummies for type of publication (book, book chapter, article), for international co-authorships, for number of authors (2 to 5, and more than 5), for papers written in Italian, and for papers evaluated by bibliometric analysis. Books, book chapters, and other research outputs (e.g. designs, architectural plans, databases, software) receive a lower evaluation relative to the excluded category (journal articles). We control also for two variables that proxy for the ability or willingness to engage in networking activities - number of authors per research paper, and presence of an international co-author. The quality score is 14 percentage points higher for papers that are internationally co-authored, and increases with the number of co-authors. Papers written in Italian receive a score that is 20 percentage points lower than the score for papers written in English or some other language than Italian. This most likely reflects the fact that publication in Italian means that the research results will be less widely disseminated than if they were published in English; usually, in many research areas only less valuable results are published in Italian journals. Finally, the coefficient of the dummy for female does not change in magnitude and significance.

Table 4 column (3) includes as a control a dummy for the evaluation method and shows that bibliometric evaluation is associated with a score that is some 20 percentage points higher compared to peer review evaluation.[11] The coefficient of the gender gap does not change appreciably. In the last specification (column 4), we add as a control the position in the institution of the researcher submitting the scientific output (associate professor and full professor, and equivalent positions in a research institutes). The category of assistant professor is excluded. The most interesting result is that the difference in research quality between men and women is only 1% but still significant at the 1% level.

[11] This difference is due partly to a quality effect. For instance, in scientific areas such medicine, chemistry, and physics, papers not published in journals which often are less original and of lower impact, were evaluated by peer review. The second and more important effect is that in our data, bibliometric evaluation tends to be more generous than peer review. This was highlighted by Cicero et al. (2014) who compare the two evaluations in a random sample for which both evaluations were available using the same dataset of the VQR 2004-2010.

In Table 5 the sample is split by academic rank of the research staff. The gender gap is relatively small in all three subsamples (1% or less) but in all cases is statistically different from zero at the 1% level. Overall, Tables 3 and 4 show that the gender gap is only slightly affected by the observable characteristics of the papers and the evaluation method, and the gap narrows significantly when we control for the author's academic position.

The data description in Table 2 shows that most of the gender gap is due to a lower probability of female authored papers obtaining an excellent evaluation. Table 6 presents the results of a linear modeling of the probability of obtaining an excellent score, controlling for all the above mentioned observables.

In the baseline regression being female reduces the probability of obtaining an excellent evaluation by 6 percentage points (column 1). If we include the characteristics of the paper and of the evaluation method, the magnitude of the female dummy drops to 5 points, and then to 3 points if we control also for academic rank. Estimating the regression by probit yields similar results. In Table 7 we repeat the estimation splitting the sample according to academic rank. Regardless of the position held, the probability that women obtain top evaluations is lower than for men, with values ranging from -3.3% for full professors, -2.2% for associate professors, and -2.5% for assistant professors. These results are qualitatively similar to those obtained in the score analysis: the gender gap falls substantially once we control for academic position, regardless of the position held.

Table 8 presents a replication of the most complete specification with separate regressions run for each research area, and quality score and a dummy for top papers as the dependent variables. To save space, we report only the coefficients of interest. We find that the dummy for female is negative and statistically different from zero for 10 out 14 areas, and not statistically different from zero in two areas (agricultural and veterinary sciences, and psychology) in both regressions. The biggest gaps are in mathematics and civil engineering, where the women's quality score is some 3 percentage points lower than the men's score, and the probability of a top evaluation is 5 to 6 points lower for women, even controlling for all observable characteristics of papers and researchers (including academic position).

In the succeeding sections we explore the reasons for the existence of the gender gap in research, and check whether the gap is larger for women who have taken maternity leave, or if it is due to discrimination by referees, comparing bibliometric evaluation and peer review for

a random sample of papers that were evaluated using both methods. We check also whether the evaluation method amplifies the gender gap.

5. The effect of family responsibilities on women's research quality

There is no agreement in the literature about the role of family responsibilities and child-rearing on women's scientific production and careers. Some studies find that motherhood does not play a relevant role in gender differences in scientific productivity (Fox, 2005; Stack, 2004; Krapf et al., 2014) or finds a positive relationship between fertility and academic output (Joecks et al., Pull, and Backes-Gellner, 2014). Other studies identify motherhood choice and engagement in child care as prominent reasons for the underrepresentation of women in science (Ceci et al., 2011).

The literature provides little evidence of the effect of motherhood on research quality, probably due to lack of data. A recent paper concludes that women who have experienced maternity leave receive lower evaluations for their papers assessed based on journal metrics and journal ratings (Brooks et al., 2014).

We can explore this question in the context of research quality by merging our dataset of papers and researchers with data on periods of leave (maternity, health reasons, research, etc.) provided by the Ministry of Education, Universities and Research (MIUR).[12] Provision of data by universities and public research centers is voluntary and therefore may not include all researchers' leave. The MIUR dataset includes some 24,000 leave periods between 1973 and 2010. The number of observations pre-1990 is smaller, in part because the data were not collected, and in part because the proportion of women in academia has changed over time. Maternity leave accounts for one-third of all leave time reported in the dataset (34%).

During the five-month period of compulsory leave from work women receive a maternity allowance in lieu of pay. After five months, they can take voluntary leave at reduced pay. Given the discretionary nature of this leave, in this section we focus only on compulsory maternity leave taken before the evaluation period (2004-2010). We do not consider maternity leave in 2004-10 because the VQR call makes allowances for women who experienced maternity leave during this period; e.g. researchers who had a child in 2004-10 are required to submit two papers. Since there is a potential endogeneity problem between

[12] We thank MIUR for providing these data. The merging with our dataset was anonymized.

discretionary maternity leave and research performance, we focus only on compulsory leave periods, checking whether having one or more children before 2004 affects the woman's research performance in 2004-2010, compared to either women without children, or men.

The regressions in Table 9 include the number of days of compulsory maternity leave taken before 2004, and report the results for quality score and the probability of an excellent evaluation, in the total sample, and in subsamples of assistant, associate, and full professor. In the regressions for quality score the sign of the maternity leave coefficient is always negative but is imprecisely estimated; the coefficient is statistically different from zero at the 10% level only in the regression for assistant professors. In the regressions for the probability of a paper being evaluated as excellent the maternity leave coefficient is negative, and statistically different from zero in the total sample and in the sample of associate professors. Overall, the results in Table 9 show that women who experienced maternity leave before 2004-2010, tend to receive lower evaluations for papers written during that period. However, regardless of the sign and significance of the maternity leave coefficient, controlling for leave does not affect the coefficient of gender in any of the regressions, and therefore does not affect the gender gap in research.[13]

6. Referee's gender

In this section we analyze how the presence of women among referees affects the final evaluation of the quality of the research output. In other words, we want to explore whether males tend to discriminate towards women, and whether women tend to write more favorable evaluations for papers authored by females.

The effect of evaluators' gender has been studied in relation to grant awards (Broder 1993) and academic promotion (Bagues et al. 2014; De Paola and Scoppa, 2015). The empirical evidence does not offer conclusive evidence on discrimination. Some studies find that researchers benefit from the presence of same gender evaluators (De Paola and Scoppa 2015); others find an opposite-sex preference among evaluators (Broder, 1993; Bagues et al. 2014), and yet others find no significant role of gender (Zinovyeva and Bagues, 2011)).

[13] We tried different specifications using as alternative regressors the number of compulsory maternity leave periods and a dummy for maternity leave. None of these coefficients is statistically different from zero.

We are able to address this issue using data on referees' reports and referees' characteristics – gender, age, affiliation - which we merge with the initial dataset.[14] In the VQR, peer review evaluation is organized as follows. Panel members assign each paper to two external referees chosen independently by two experts on the panel. The referee report is organized around three sections (originality, relevance, and international outreach) scored by the referee in the range 3 to 27. The two referee reports are then averaged to obtain a single score which is converted by the panel into a final merit class (0, 0.5, 0.8, and 1).

Table 10 reports two different specifications - one where the dependent variables are the final quality score (column 1), and one which includes a dummy variable for papers evaluated as excellent (column 2). Each specification uses the same variables as in our baseline regressions plus variables for the sum of the ages of the two referees, whether both referees are affiliated to an Italian institution, and dummies for the gender composition of the referee committee. We define dummy variables for if the paper was evaluated by two female referees, and if one of the two referees is female, and include an interaction term between gender of the researcher submitting the research output, and dummies for the gender composition of the referees.

The results suggest that on average, females give more generous evaluations. In column 1, the average score is 2.4 percentage points higher if both referees are females, and 1.7 percentage points higher if one of the two referees is female, with respect to papers evaluated by two male referees. Column 2 shows there is no evidence of an association between referee's gender and the probability of the paper being assessed as excellent. We found evidence of a *same sex preference* if a paper submitted by a woman is refereed by two women: the coefficient of the interaction terms "*Female* × *Both referees are female*" is positive and statistically different from zero at the 1% level if the dependent variable is quality score (column 1).[15]

[14] The merge was totally anonymized.

[15] The results reported in Table 10 were obtained dropping from the sample all papers evaluated by bibliometric analysis. We replicated the regressions including only research areas where all papers were evaluated by peer review and the results were similar.

7. Bibliometric evaluation vs. peer review

There may be several reasons for the gender gap in research evaluation. On the one hand, as highlight in Section 6, there is some evidence of a for "same sex preference" in evaluations. Since in VQR 2004-2010 the majority of reviewers were males (around 11,ooo, or 74% of the total number of referees), this might bias peer review evaluation against females. On the other hand, to the extent that there is a "same sex preference for citations", and that in many areas the majority of scientists are male (see Table 3), bibliometric evaluation might be lower for women.

Previous studies find that women's papers attract fewer citations than men's (Ward et al., 1992; Davenport and Snyder, 1995) even when controlling for the characteristics of the paper and the researcher(s) (HEFCE, 2011). For instance, Larivière et al. (2013) found that when a woman has a prominent authorship position (sole author, first author, last author), the paper receives fewer citations compared to the same parameters for men. There is evidence also that the use of journal metrics to score papers penalizes women (Brooks et al. 2014).

However, the studies cited do not compare peer review and bibliometric evaluation in order to understand whether peer review might correct (or worsen) the supposed disadvantage that bibliometric evaluation imposes on women. We were able to make this comparison using VQR data. A distinctive and very useful feature of VQR data is that, for statistical purposes, a random sample of 10% of all papers evaluated by bibliometric analysis were also evaluated by peer review. The sample of nearly 7,500 papers was stratified by research areas, and includes all scientific sectors for which bibliometric indices are available and reliable for part or all of the papers (i.e. it does not include humanities, law, or sociology). It is important to note that the final evaluation of these papers was based on the bibliometric indicators, and that peer review reports were collected only for statistical purposes. Indeed, the random sample allows a thorough statistical comparison between the two evaluation methods, and in particular, the degree of agreement between bibliometric evaluation and peer review.[16]

Table 11 reports the regressions for quality score and for the probability of the paper being evaluated as excellent, separately for the two evaluation methods. Columns (1) and (3) provide the peer review assessments, and columns (2) and (4) provide the evaluation based on

[16] Cicero et al. (2015) report detailed statistics by research area on the difference between the two types of evaluations. Bertocchi et al. (2015) compare papers published in economics, management, and statistics.

a combination of journal impact factor and citations received by individual papers. The most notable pattern in Table 11 is that bibliometric evaluations tend to be more generous than peer review evaluations. This is evidenced by the values of the constant terms in columns (1) and (2) which are considerably higher if the papers are evaluated by bibliometric indicators.

In the context of the present paper, the coefficient of the female dummy is interesting. In column (1), for papers evaluated by peer review, the coefficient is -0.031 and is precisely estimated but close to zero for papers evaluated by bibliometric indicators. In columns (3) and (4), the probability of an excellent evaluation is around 3 percentage points lower for both bibliometric and peer review evaluation. Overall, we find some evidence that bibliometric evaluation tends to be slightly more favorable for women than peer review.[17]

8. Allocation and submission of papers

In this section we investigate whether a gender gap is due to the allocation of papers with multiple authors to individual researchers. According to the VQR rules, each researcher submits three papers for evaluation, and coauthored papers can be submitted only by one researcher in each institution. If the institution fails to follow this rule, the paper is excluded from the evaluation; that is, multiple submissions of the same paper by the same institution are not allowed. Therefore, in each university or research center, each paper is associated with only one researcher affiliated to that institution, identified as "the author of the paper".

To fulfill this requirement, institutions, in turn, ask their research staff to prepare a list of papers in a number typically exceeding three, ordering the papers by self-assessed quality in order to maximize their final score. When the same paper appears more than once, the department chair (or a delegate) allocates the paper to only one researcher. During the process, discrimination may operate precisely at this allocation stage: in case of multiple submissions of coauthored papers, and in case the author is a woman, the institutions may allocate their best papers (for instance, those published in highly cited journals) to men instead of women.

To investigate this issue, we replicate the baseline regressions focusing only on the set of almost 50,000 single-authored papers. If the gender gap attenuates or disappears for this

[17] We also ran fixed effects estimates pooling together peer and bibliometric evaluations and the results were confirmed.

sample, then one could infer that discrimination operates at the selection and allocation stages of papers. Otherwise, if results are similar to the whole sample, we should conclude that the gap does not arise from this particular form of discrimination.

We replicate the baseline specifications, using both the quality score and the fraction of excellent papers as dependent variables, as well as regressions by academic rank using the sample of single-authored papers. It should be noticed, that this selected sample includes mostly papers in Humanities, Law and Social Sciences, where the fraction of single author papers is much larger than in other research areas.

Table 12 reports results for the determinants of quality score and the probability that the paper receives a top evaluation for single author papers using the most complete specification. In column 1 the female dummy is slightly less than 1 percentage point and statistically significant at the 5% level, Therefore results are similar to those found in Table 4 (column 4). In column 2 we find that the gender gap persists for top evaluations, but again there are no major differences with respect to the full sample estimates. In particular the probability that single-authored papers by women receive a top evaluation is 2.4 percentage points lower than for men, against a gap of 2.7 percentage points in the full sample estimates (column 4 of Table 6). We also run regressions splitting the sample by academic rank, and find similar patterns as in the full sample estimates. For brevity, these results are not reported and are available on request. We conclude from this analysis that no discrimination arises at the allocation stage of papers with multiple authors to individual researchers.

9. Conclusions

This paper contributes to the literature on the gender gap in research. We exploited a large dataset of some 180,000 papers evaluated for the 2004-2010 assessment of Italian universities and research institutions. The dataset provides detailed information on type publication, evaluation method (peer review or bibliometric analysis), and the characteristics of authors and referees. To measure research quality, we associated to each paper; quality score (ranging from 0 to 1), and a dummy equal to 1 for papers classified by bibliometric analysis or by referees as “excellent” - thus receiving a top evaluation.

In our baseline estimate, we found evidence of a gender gap in research equal to 5 percentage points for the quality score and 7 percentage points for a top evaluation.

Controlling for research output characteristics and university rank (assistant, associate, full professor or equivalent), gender inequality falls sharply (to 1%) but the gap persists in the regressions for the probability of receiving the highest score (2 to 3 percentage points). In the rest of the paper we checked several dimensions that might explain the gap, focusing on external constraints (family commitments) and discrimination, to understand whether the gap is larger for women who experienced maternity leave, or if it stems from discrimination. In the case of family duties, our results suggest that maternity leave does not play a major role in explaining the gap. Identifying the presence of discrimination is difficult since it can take many forms. We explored two potential sources: discrimination against women by referees, and discrimination against women by bibliometric evaluation. We analyzed the sample of peer reviewed papers and found some evidence of a "same sex preference" although the gap was unaffected. We explored the random sample of journal articles evaluated by both peer review and bibliometric evaluation: comparison between the two methodologies reveals that bibliometric evaluations tend to be more generous than peer review but that for both the gender gap persists – especially for top evaluations. Finally we replicate the same analysis by focusing on a sample of single-authored papers. This robustness check reveals that no discrimination arises before the evaluation, when institutions allocate papers with multiple authors to individuals.

Overall, we found no evidence that the VQR is "unfair" to women. Our finding of a persistent (albeit small) gap for top papers might be due to a "selection effect" rather than an "evaluation effect": women tend to choose slightly less risky projects and research strategies which results in a lower fraction of top papers. This is consistent with the literature on gender differences in preferences which suggests that women are more risk averse than men (Borghans et al., 2009) and less competitive (Niederle and Vesterlund, 2005).

References

Ancaiani A. et al. (2015), "Evaluating scientific research in Italy: The 2004-10 research evaluation exercise", *Research Evaluation* 5; DOI:10.1093/reseval/rvv008

Bagues M., Sylos-Labini M. and Zinovyeva N. (2014), "Do gender quotas pass the test? Evidence from academic evaluations in Italy", LEM Working Paper Series 2014/14. http://dx.doi.org/10.2139/ssrn.2457487

Bertocchi, G., Gambardella, A., Jappelli, T. Nappi, C.A., Peracchi, F. (2015), "Bibliometric evaluation vs. informed peer review: Evidence from Italy", *Research Policy* 44, 451-466.

Borghans, L., Heckman, J. J., Golsteyn, B. H., & Meijers, H. (2009), "Gender differences in risk aversion and ambiguity aversion", *Journal of the European Economic Association*, 7, 649-658.

Brooks, C., Fenton E.M., Walker J.T. (2014), "Gender and the evaluation of research", *Research Policy* 43, 990-1001.

Ceci, S. J., & Williams, W. M. (2011), "Understanding current causes of women's underrepresentation in science", *Proceedings of the National Academy of Sciences*, 108, 3157-3162.

Cicero, T., Malgarini, M., Nappi, C. A. (2013), "Bibliometric and peer review methods for research evaluation: a methodological appraisement", MPRA (Munich Personal REPEC Archive).

Davenport, E., Snyder, H., (1995), "Who cites women? Whom do women cite? An exploration of gender and scholarly citation in sociology", *Journal of Documentation* 51, 404–410

De Paola M., Ponzo M., Scoppa V. (2015), "Gender differences in attitudes towards competition: Evidence from the Italian scientific qualification", CSEF Working Papers No. 391, http://www.csef.it/WP/wp391.pdf

De Paola, M., Scoppa, V. (2015), "Gender discrimination and evaluators' gender: Evidence from Italian academia", *Economica* 82, 162-188.

Fox, M. (2005), "Gender, family characteristics, and publication productivity among scientists", *Social Studies of Science* 35, 131–150.

Higher Education Founding Council for England (2011), "Analysis of data from the pilot exercise to develop bibliometric indicators for the REF. The effect of using normalised citation scores for particular staff characteristics", HEFCE Issue paper, February 2011/03

Kaufman, G., Uhlenberg, P. (2000), "The influence of parenthood on the work effort of married men and women", *Social Forces* 78, 931–947.

Krapf M. , Ursprung, H.W., Zimmermann, C. (2014), "Parenthood and productivity of highly skilled labor: Evidence from the groves of academe", IZA Discussion Paper No. 7904.

Larivière V., Ni C., Gingras Y., Cronin B., Sugimoto C.R. (2013), "Global gender disparities in science", *Nature* 504.

Larivière V., Vignola-Gagné E., Villeneuve C., Gélinas P., Gingras Y. (2011), "Sex differences in research funding, productivity and impact: an analysis of Quebec university professors", *Scientometrics* 87, 483-498.

Ley T.J. and Hamilton B.H. (2008), "The gender gap in NIH grant applications", *Science* 322, 1472-1474.

Leslie S.J., Cimpian A., Meyer M., Freeland E. (2015), "Expectations of brilliance underlie gender distributions across academic disciplines", *Science* 347, 262-265.

Mauleón M. and Bordons (2006), "Productivity, impact and publication habits by gender in the area of Materials Science", *Scientometrics* 66, 199–218.

Niederle, M., Vesterlund, L. (2005), "Do women shy away from competition?", NBER Working Paper No. 11474.

Stack, S. (2004), "Gender, children and research productivity", *Research in Higher Education* 45, 891–920.

Joecks, J., Pull, K., Backes-Gellner U. (2014), "Childbearing and (female) research productivity - a personnel economics perspective on the leaky pipeline", *Journal of Business Economics* 84, 517-530.

Ward, K.B., Gast, J., Grant, L., (1992), "Visibility and dissemination of women's and men's sociological scholarship", *Social Problems* 39, 291–298.

West JD, Jacquet J, King MM, Correll SJ, Bergstrom CT (2013), "The role of gender in scholarly authorship", *PLoS ONE* 8: e66212. doi:10.1371/journal.pone.0066212

Zinovyeva, N., Bagues, M.(2011), "Does gender matter for academic promotion? Evidence from a randomized natural experiment", IZA Discussion Paper 5537.

Table 1. Sample statistics, by gender

	Males		*Females*		*Total*	
	Mean	s.d.	Mean	s.d.	Mean	s.d.
Quality score of the paper	0.656	0.390	0.626	0.387	0.646	0.390
Excellent paper	0.391	0.488	0.319	0.466	0.366	0.482
Age less than 40	0.117	0.321	0.157	0.364	0.131	0.337
Age 40 to 55	0.536	0.499	0.585	0.493	0.552	0.497
Age over 55	0.348	0.476	0.258	0.438	0.317	0.465
Journal article	0.751	0.433	0.706	0.455	0.736	0.441
Book	0.079	0.269	0.086	0.280	0.081	0.273
Book chapter	0.101	0.302	0.140	0.347	0.114	0.318
Other research paper	0.069	0.254	0.069	0.253	0.069	0.253
International co-authorship	0.232	0.422	0.191	0.393	0.218	0.413
Written in Italian	0.214	0.410	0.260	0.439	0.230	0.421
Number of authors: 1	0.252	0.434	0.313	0.464	0.273	0.445
Number of authors: 2 to 5	0.458	0.498	0.401	0.490	0.439	0.496
Number of authors: more than 5	0.290	0.454	0.286	0.452	0.288	0.453
Bibliometric evaluation	0.473	0.499	0.424	0.494	0.456	0.498
Full Professor (or equivalent)	0.343	0.475	0.166	0.372	0.283	0.450
Associate Professor (or equivalent)	0.322	0.467	0.316	0.465	0.320	0.466
Assistant Professor (or equivalent)						
Researcher (or equivalent)	0.334	0.472	0.518	0.500	0.397	0.489
Number of observations	118,949		61,791		180,740	

Table 2. Quality score of papers, by gender

	Males		*Females*		*Total*	
	No.	%	No.	%	No.	%
1.0: Excellent	46,457	39.1	19,701	31.9	66,158	36.6
0.8: Good	29,604	24.9	17,844	28.9	47,448	26.3
0.5: Acceptable	15,869	13.3	9,409	15.2	25,278	14.0
0.0: Limited	27,019	22.7	14,837	24.0	41,856	23.2
Total	118,949	100.0	61,791	100.0	180,740	100.0

Table 3. Research quality indicators, by gender

Research Area	*Quality score*		*Difference*	*Excellent paper*		*Difference*	*% of females*
	Males	Females		Males	Females		
Mathematics and Computer Sciences	0.728	0.643	0.085***	0.490	0.368	0.122***	33.01
Physics	0.806	0.786	0.020***	0.582	0.539	0.043***	22.06
Chemistry	0.827	0.797	0.030***	0.590	0.543	0.048***	43.57
Earth Sciences	0.567	0.539	0.028***	0.320	0.271	0.049***	31.33
Biology	0.686	0.647	0.039***	0.449	0.386	0.063***	52.42
Medicines	0.602	0.595	0.007	0.393	0.361	0.032***	29.8
Agricultural and Veterinary Sciences	0.605	0.628	-0.023**	0.422	0.449	-0.028***	36.28
Civil engineering	0.639	0.648	-0.009	0.436	0.412	0.024	17.91
Architecture	0.508	0.526	-0.018*	0.100	0.077	0.022***	35.43
Industrial and Information Engineering	0.74	0.762	-0.022***	0.526	0.541	-0.015	15.31
Humanities	0.726	0.679	0.047 ***	0.282	0.206	0.076***	54.90
History, Geography, Pedagogy	0.624	0.605	0.019***	0.176	0.136	0.039***	40.31
Psychology	0.582	0.581	0.001	0.380	0.338	0.041**	53.95
Law	0.605	0.558	0.047***	0.128	0.080	0.048***	35.48
Economic and Statistics	0.384	0.351	0.033***	0.210	0.160	0.050***	34.11
Social Sciences	0.494	0.446	0.048***	0.100	0.069	0.031***	37.68
Total	0.656	0.626	0.030***	0.391	0.319	0.072***	35.34

Note. The table reports the average quality score and the proportion of papers classified as excellent, by gender of researchers. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

Table 4. The determinants of the quality score

	(1)	(2)	(3)	(4)
Female	-0.0397***	-0.0339***	-0.0320***	-0.0107***
	(0.00185)	(0.00175)	(0.00170)	(0.00170)
Age less than 40	0.135***	0.0999***	0.0891***	0.199***
	(0.00288)	(0.00274)	(0.00266)	(0.00305)
Age 40 to 55	0.0872***	0.0673***	0.0622***	0.119***
	(0.00192)	(0.00182)	(0.00177)	(0.00192)
Book		-0.0479***	0.000531	-0.00185
		(0.00364)	(0.00357)	(0.00352)
Book chapter		-0.150***	-0.0937***	-0.0935***
		(0.00312)	(0.00309)	(0.00304)
Other research output		-0.265***	-0.181***	-0.173***
		(0.00341)	(0.00342)	(0.00337)
International co-authorship		0.139***	0.123***	0.114***
		(0.00216)	(0.00211)	(0.00208)
Written in Italian		-0.203***	-0.175***	-0.165***
		(0.00283)	(0.00277)	(0.00273)
Number of authors: 2 to 5		0.0498***	0.0388***	0.0355***
		(0.00328)	(0.00319)	(0.00315)
Number of authors: more than 5		0.110***	0.0909***	0.0886***
		(0.00385)	(0.00375)	(0.00370)
Bibliometric evaluation			0.202***	0.195***
			(0.00200)	(0.00197)
Full Professor (or equivalent)				0.163***
				(0.00229)
Associate Professor (or equivalent)				0.0720***
				(0.00193)
Constant	0.267***	0.415***	0.365***	0.234***
	(0.0123)	(0.0120)	(0.0117)	(0.0117)
Observations	180,740	180,740	180,628	180,628
R-squared	0.191	0.278	0.317	0.336

Note. The table reports OLS regressions for the quality score. Each regression includes a full set of 367 dummies for scientific sectors, and 129 dummies for universities and research institutions. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

Table 5. The determinants of the quality score, by academic position of the author

	Full Professors	Associate Professors	Assistant Professors
	(1)	(2)	(3)
Female	-0.00712**	-0.00666**	-0.0114***
	(0.00356)	(0.00296)	(0.00263)
Age less than 40	0.120***	0.218***	0.204***
	(0.0190)	(0.00679)	(0.00474)
Age 40 to 55	0.0889***	0.143***	0.134***
	(0.00300)	(0.00315)	(0.00417)
Book	-0.0115*	-0.00573	0.0129**
	(0.00591)	(0.00620)	(0.00611)
Book chapter	-0.0829***	-0.0973***	-0.101***
	(0.00511)	(0.00546)	(0.00519)
Other research output	-0.153***	-0.171***	-0.186***
	(0.00627)	(0.00600)	(0.00538)
International co-authorship	0.108***	0.108***	0.124***
	(0.00375)	(0.00361)	(0.00344)
Written in Italian	-0.146***	-0.166***	-0.172***
	(0.00477)	(0.00485)	(0.00459)
Number of authors: 2 to 5	0.0417***	0.0338***	0.0408***
	(0.00555)	(0.00541)	(0.00536)
Number of authors: more than 5	0.0780***	0.0939***	0.0954***
	(0.00680)	(0.00640)	(0.00611)
Bibliometric evaluation	0.191***	0.194***	0.194***
	(0.00365)	(0.00346)	(0.00316)
Constant	0.337***	0.315***	0.245***
	(0.0200)	(0.0187)	(0.0213)
Observations	51,057	57,812	71,759
R-squared	0.344	0.341	0.337

Note. The table reports OLS regressions for the quality score. Each regression includes a full set of 367 dummies for scientific sectors, and 129 dummies for universities and research institutions. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

Table 6. Determinants of the probability that the paper is excellent

	(1)	(2)	(3)	(4)
Female	-0.0572***	-0.0510***	-0.0476***	-0.0270***
	(0.00229)	(0.00221)	(0.00209)	(0.00211)
Age less than 40	0.136***	0.102***	0.0835***	0.188***
	(0.00356)	(0.00345)	(0.00327)	(0.00377)
Age 40 to 55	0.0810***	0.0621***	0.0533***	0.109***
	(0.00237)	(0.00230)	(0.00218)	(0.00238)
Book		-0.0690***	0.0152***	0.0130***
		(0.00459)	(0.00439)	(0.00435)
Book chapter		-0.147***	-0.0485***	-0.0484***
		(0.00394)	(0.00379)	(0.00376)
Other research output		-0.267***	-0.121***	-0.113***
		(0.00430)	(0.00420)	(0.00417)
International co-authorship		0.183***	0.155***	0.146***
		(0.00272)	(0.00259)	(0.00257)
Written in Italian		-0.126***	-0.0788***	-0.0688***
		(0.00358)	(0.00340)	(0.00338)
Number of authors: 2 to 5		0.0378***	0.0186***	0.0152***
		(0.00414)	(0.00392)	(0.00389)
Number of authors: more than 5		0.105***	0.0706***	0.0678***
		(0.00486)	(0.00461)	(0.00457)
Bibliometric evaluation			0.352***	0.345***
			(0.00245)	(0.00244)
Full Professor (or equivalent)				0.159***
				(0.00283)
Associate Professor (or equivalent)				0.0616***
				(0.00239)
Constant	0.0232	0.131***	0.0449***	-0.0796***
	(0.0151)	(0.0152)	(0.0144)	(0.0145)
Observations	180,740	180,740	180,628	180,628
R-squared	0.192	0.248	0.325	0.337

Note: The table reports the results of a linear probability model where the dependent variable is whether the paper is excellent. Each regression includes a full set of 367 dummies for scientific sectors, and 129dummies for universities and research institutions. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

Table 7. Determinants of the probability that the paper is excellent, by academic position of the author

	Full Professors	Associate Professors	Assistant Professors
	(1)	(2)	(3)
Female	-0.0333***	-0.0222***	-0.0246***
	(0.00480)	(0.00367)	(0.00307)
Age less than 40	0.150***	0.221***	0.160***
	(0.0255)	(0.00840)	(0.00554)
Age 40 to 55	0.106***	0.128***	0.0872***
	(0.00405)	(0.00390)	(0.00488)
Book	0.0159**	0.0137*	0.0146**
	(0.00796)	(0.00768)	(0.00715)
Book chapter	-0.0532***	-0.0532***	-0.0410***
	(0.00689)	(0.00676)	(0.00607)
Other research output	-0.119***	-0.110***	-0.111***
	(0.00845)	(0.00742)	(0.00628)
International co-authorship	0.143***	0.142***	0.150***
	(0.00505)	(0.00447)	(0.00402)
Written in Italian	-0.0893***	-0.0643***	-0.0535***
	(0.00642)	(0.00600)	(0.00537)
Number of authors: 2 to 5	0.0173**	0.0180***	0.0165***
	(0.00748)	(0.00669)	(0.00626)
Number of authors: more than 5	0.0531***	0.0748***	0.0746***
	(0.00916)	(0.00792)	(0.00714)
Bibliometric evaluation	0.351***	0.352***	0.334***
	(0.00492)	(0.00428)	(0.00370)
Constant	0.0596**	-0.0312	-0.0441*
	(0.0269)	(0.0232)	(0.0248)
Observations	51,057	57,812	71,759
R-squared	0.335	0.352	0.338

Note. The table reports the results of a linear probability model where the dependent variable is whether the paper is excellent. Each regression includes a full set of 367 dummies for scientific sectors, and 129 dummies for universities and research institutions. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

Table 8: Regressions by research area

	Quality score		*Paper is excellent*	
	Coefficient of female dummy	s.e	Coefficient of female dummy	s.e
Mathematics and Computer Sciences	-0.0378***	(0.00697)	-0.0663***	(0.00931)
Physics	-0.00468	(0.00492)	-0.0188**	(0.00763)
Chemistry	-0.0210***	(0.00569)	-0.0284***	(0.00916)
Earth Sciences	-0.0223**	(0.00865)	-0.0402***	(0.0102)
Biology	-0.00665	(0.00588)	-0.0234***	(0.00741)
Medicines	-0.0132**	(0.00526)	-0.0276***	(0.00623)
Agricultural and Veterinary Sciences	-0.00165	(0.00761)	0.00113	(0.00872)
Civil engineering	-0.0299**	(0.0143)	-0.0468***	(0.0168)
Architecture	0.0199**	(0.01000)	-0.0203**	(0.00836)
Industrial and Information Engineering	0.00675	(0.00625)	-0.00442	(0.00869)
Humanities	-0.0117**	(0.00510)	-0.0378***	(0.00740)
History, Geography, Pedagogy	0.00683	(0.00697)	-0.00599	(0.00777)
Psychology	0.0129	(0.0117)	-0.0189	(0.0130)
Law	-0.00801	(0.00623)	-0.0179***	(0.00621)
Economic and Statistics	-0.0135**	(0.00637)	-0.0254***	(0.00675)
Social Sciences	-0.0300***	(0.0105)	-0.0152*	(0.00899)

Note. For each research area, the table reports separate OLS regressions for quality score, and linear probability models where the dependent variable is whether the paper is excellent. Each regression includes a full set of 129 dummies for universities and research institutions. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

Table 9. The effect of maternity leaves on research quality

	All sample		Full Professors		Associate Professors		Researchers	
	Quality score	Paper is excellent	Quality score	Paper is excellent	Quality score	Paper is excellent	Quality score	Paper is excellent
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-0.0102***	-0.0261***	-0.00694*	-0.0325***	-0.00644**	-0.0201***	-0.0107***	-0.0244***
	(0.00173)	(0.00214)	(0.00363)	(0.00489)	(0.00303)	(0.00375)	(0.00265)	(0.00310)
Age less than 40	0.198***	0.188***	0.120***	0.150***	0.218***	0.221***	0.204***	0.160***
	(0.00305)	(0.00377)	(0.0190)	(0.0255)	(0.00679)	(0.00840)	(0.00475)	(0.00555)
Age 40 to 55	0.119***	0.110***	0.0889***	0.106***	0.143***	0.128***	0.134***	0.0872***
	(0.00193)	(0.00238)	(0.00301)	(0.00405)	(0.00316)	(0.00391)	(0.00417)	(0.00488)
Book	-0.00186	0.0130***	-0.0115*	0.0159**	-0.00574	0.0136*	0.0129**	0.0146**
	(0.00352)	(0.00435)	(0.00591)	(0.00796)	(0.00620)	(0.00768)	(0.00611)	(0.00715)
Book chapter	-0.0935***	-0.0485***	-0.0829***	-0.0532***	-0.0973***	-0.0533***	-0.101***	-0.0410***
	(0.00304)	(0.00376)	(0.00511)	(0.00689)	(0.00546)	(0.00676)	(0.00519)	(0.00607)
Other research output	-0.173***	-0.113***	-0.153***	-0.119***	-0.171***	-0.110***	-0.186***	-0.111***
	(0.00337)	(0.00417)	(0.00627)	(0.00845)	(0.00600)	(0.00742)	(0.00538)	(0.00628)
International co-authorship	0.114***	0.146***	0.108***	0.143***	0.108***	0.142***	0.124***	0.150***
	(0.00208)	(0.00257)	(0.00375)	(0.00505)	(0.00361)	(0.00447)	(0.00344)	(0.00402)
Research output in Italian	-0.165***	-0.0688***	-0.146***	-0.0893***	-0.166***	-0.0643***	-0.172***	-0.0535***
	(0.00273)	(0.00338)	(0.00477)	(0.00642)	(0.00485)	(0.00600)	(0.00459)	(0.00537)
Number of authors: 2 to 5	0.0356***	0.0152***	0.0417***	0.0173**	0.0339***	0.0181***	0.0409***	0.0165***
	(0.00315)	(0.00389)	(0.00555)	(0.00748)	(0.00541)	(0.00669)	(0.00536)	(0.00626)
Number of authors: more than 5	0.0886***	0.0679***	0.0780***	0.0532***	0.0939***	0.0749***	0.0955***	0.0747***
	(0.00370)	(0.00457)	(0.00680)	(0.00916)	(0.00640)	(0.00792)	(0.00611)	(0.00714)
Bibliometric evaluation	0.195***	0.345***	0.191***	0.351***	0.194***	0.352***	0.194***	0.334***
	(0.00197)	(0.00244)	(0.00365)	(0.00492)	(0.00346)	(0.00428)	(0.00316)	(0.00370)
Days of compulsory maternity leave before 2004	-0.0431	-0.0817**	-0.0150	-0.0706	-0.0144	-0.138***	-0.0923*	-0.0361
	(0.0278)	(0.0344)	(0.0624)	(0.0841)	(0.0418)	(0.0518)	(0.0474)	(0.0555)
Full Professor (or equivalent)	0.163***	0.159***						
	(0.00229)	(0.00283)						
Associate Professor (or equivalent)	0.0721***	0.0617***						
	(0.00193)	(0.00239)						
Constant	0.234***	-0.0799***	0.337***	0.0595**	0.315***	-0.0318	0.245***	-0.0441*
	(0.0117)	(0.0145)	(0.0200)	(0.0269)	(0.0187)	(0.0232)	(0.0213)	(0.0248)
Observations	180,628	180,628	51,057	51,057	57,812	57,812	71,759	71,759
R-squared	0.336	0.337	0.344	0.335	0.341	0.352	0.337	0.338

Note. The table reports OLS regressions for the quality score and linear probability models where the dependent variable is whether the paper is excellent. Each regression includes a full set of 367 dummies for scientific sectors and 129 dummies for universities and research institutions. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

Table 10. The effect of referees' gender

	Quality score	Paper is excellent
	(1)	(2)
Female	-0.00652**	-0.0196***
	(0.00300)	(0.00277)
Age less than 40	0.182***	0.0906***
	(0.00421)	(0.00389)
Age 40 to 55	0.113***	0.0598***
	(0.00255)	(0.00236)
Book	0.00940***	0.0216***
	(0.00362)	(0.00334)
Book chapter	-0.0759***	-0.0344***
	(0.00316)	(0.00291)
Other research output	-0.136***	-0.0503***
	(0.00353)	(0.00326)
International co-authorship	0.119***	0.0657***
	(0.00365)	(0.00337)
Written in Italian	-0.129***	-0.0414***
	(0.00298)	(0.00275)
Number of authors: 2 to 5	0.0272***	0.0105***
	(0.00372)	(0.00343)
Number of authors: more than 5	0.0768***	0.0216***
	(0.00500)	(0.00462)
Full Professor (or equivalent)	0.170***	0.105***
	(0.00309)	(0.00285)
Associate Professor (or equivalent)	0.0684***	0.0312***
	(0.00264)	(0.00244)
Sum of age the two referees	0.000299***	-2.08e-05
	(6.53e-05)	(6.03e-05)
Both referees are affiliated to an Italian institution	-0.0302***	-0.0165***
	(0.00257)	(0.00238)
Both referees are females	0.0239***	0.00771
	(0.00566)	(0.00523)
One referee is female	0.0167***	0.000988
	(0.00299)	(0.00276)
Female × both referees are female	0.0284***	0.00685
	(0.00766)	(0.00708)
Female × one referee is female	-0.00418	-0.00113
	(0.00470)	(0.00434)
Constant	0.271***	-0.00328
	(0.0245)	(0.0227)
Observations	97,576	97,576
R-squared	0.289	0.110

Note. The sample includes only papers evaluated by peer review. The table reports regressions for average quality score, sum of the scores of the two referees, and the probability of the paper being assessed as excellent. Each regression includes a full set of 367 dummies for scientific sectors, and 129 dummies for universities and research institutions. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

Table 11. The effect of the evaluation method

	Quality score by:		Paper is excellent according to:	
	Peer review	Bibliometric evaluation	Peer review	Bibliometric evaluation
	(1)	(2)	(3)	(4)
Female	-0.0301***	-0.00798	-0.0377***	-0.0312**
	(0.0101)	(0.00885)	(0.0119)	(0.0123)
Age less than 40	0.169***	0.224***	0.145***	0.284***
	(0.0149)	(0.0154)	(0.0176)	(0.0214)
Age 40 to 55	0.0992***	0.138***	0.0826***	0.162***
	(0.00979)	(0.0102)	(0.0115)	(0.0141)
Book chapter	0.00364	-0.230	-0.0141	-0.189
	(0.159)	(0.166)	(0.188)	(0.230)
Other research output	-0.532**	-0.738***	-0.0794	-0.602*
	(0.221)	(0.230)	(0.260)	(0.319)
International co-authorship	0.114***	0.104***	0.118***	0.154***
	(0.00849)	(0.00880)	(0.0100)	(0.0122)
Written in Italian	-0.325***	-0.334***	-0.0418	-0.228***
	(0.0328)	(0.0337)	(0.0386)	(0.0468)
Number of authors: 2 to 5	0.0392**	0.0704***	-0.00394	0.0352
	(0.0189)	(0.0195)	(0.0222)	(0.0271)
Number of authors: more than 5	0.0814***	0.0977***	0.0277	0.0753**
	(0.0206)	(0.0213)	(0.0243)	(0.0296)
Full Professor (or equivalent)	0.120***	0.164***	0.0935***	0.211***
	(0.0114)	(0.0118)	(0.0134)	(0.0163)
Associate Professor (or equivalent)	0.0480***	0.0771***	0.0233**	0.0782***
	(0.00942)	(0.00976)	(0.0111)	(0.0135)
Sum of age the two referees	-0.000354		-0.000689**	
	(0.000287)		(0.000338)	
Both referees are affiliated to an Italian institution	-0.0223**		-0.00660	
	(0.00881)		(0.0104)	
Both referees are females	0.0339		-0.0148	
	(0.0310)		(0.0365)	
One referee is female	0.0147		0.0115	
	(0.0108)		(0.0127)	
Female x “Both referees are females”	0.0411		-0.0122	
	(0.0454)		(0.0534)	
Female x “One referee is female”	0.0155		0.00485	
	(0.0181)		(0.0213)	
Constant	0.456***	0.378***	-0.0810	0.137
	(0.0960)	(0.0949)	(0.113)	(0.132)
Observations	7,407	7,453	7,407	7,453
R-squared	0.249	0.259	0.146	0.221

Note. The sample includes only papers evaluated by both peer review and bibliometric analysis. The table reports separate regressions for average quality score and the probability that the paper is assessed as excellent. Each regression includes a full set of 367 dummies for scientific sectors, and 129 dummies for universities and research institutions. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

Table 12. The determinants of the quality score and of the probability that the paper is excellent: single-authored papers

	(1)	(2)
Female	-0.00721**	-0.0246***
	(0.00303)	(0.00332)
Age less than 40	0.203***	0.121***
	(0.00563)	(0.00617)
Age 40 to 55	0.130***	0.0831***
	(0.00343)	(0.00376)
Book	0.0510***	0.0422***
	(0.00407)	(0.00446)
Book chapter	-0.0280***	-0.0221***
	(0.00375)	(0.00411)
Other research output	-0.0477***	-0.0287***
	(0.00532)	(0.00583)
Written in Italian	-0.0965***	-0.0753***
	(0.00362)	(0.00397)
Bibliometric evaluation	0.220***	0.244***
	(0.00757)	(0.00829)
Full Professor (or equivalent)	0.207***	0.159***
	(0.00410)	(0.00449)
Associate Professor (or equivalent)	0.0894***	0.0504***
	(0.00364)	(0.00398)
Constant	0.251***	0.0411
	(0.0689)	(0.0755)
Observations	49,299	49,299
R-squared	0.310	0.157

Note. The table reports OLS regressions for the quality score in column (1) and the results of a linear probability model where the dependent variable is whether the paper is excellent in column (2). Each regression includes a full set of 367 dummies for scientific sectors, and 129 dummies for universities and research institutions. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.