Research Policy 46 (2017) 911–924

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Gender effects in research evaluation

Tullio Jappelli[a,*], Carmela Anna Nappi[b], Roberto Torrini[c]

[a] University of Naples Federico II, Italy
[b] Anvur, Italy
[c] Bank of Italy, Italy

ARTICLE INFO

Article history:
Received 30 October 2015
Received in revised form 28 February 2017
Accepted 7 March 2017
Available online 30 March 2017

Keywords:
Research evaluation
Gender gap
Bibliometric analysis
Peer review
Maternity leaves

ABSTRACT

The paper contributes to the literature on gender gap in research investigating whether there is a gender gap in research evaluation. We use detailed data on 180,000 research papers evaluated during the Italian national research assessment (VQR 2004–2010) conducted by the Agency for the Evaluation of Universities and Research Institutes (Anvur). The data are merged with information on individual researchers and characteristics of referees. The most important empirical finding is that there is a significant gender gap in research evaluation. The gap is reduced once we control for researchers' characteristics, such as age and academic rank, but is almost unaffected by the characteristics of the research output (monographs, journal articles, book chapters, etc.), co-authorships, international collaborations. Childbearing and maternity leaves do not account for the remaining gap in research evaluation. The evaluation method (peer review or bibliometric analysis) and the referee mix (whether men or women) do not disadvantage women. Analysis of a random sample of papers evaluated using bibliometric indicators and peer review reveals that bibliometric evaluation proves to be more favourable to women than peer review evaluation.

1. Introduction

Gender gaps in the labour 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, labour economists observe persistent gaps in labour market participation and wages. Academic and research profession is no exception. Academic rankings show a persistent gender gap, particularly large in scientific fields, with a consistent pattern across different countries (European Commission, She Figures 2015 and OECD, Education at a Glance 2015). Although the gap might be narrowing over time (Ceci et al., 2014), glass ceiling is a clear concern for the research profession.

A large literature has investigated gender differences in research and academic career, studying factors affecting the career opportunity of women with respect to men: research productivity; discrimination in peer reviewing of papers, citation patterns, grant allocations and hiring practices; genetic characteristics that could affect the success opportunities in some scientific fields; preferences and family responsibility affecting time allocation and career choice and productivity. Ceci et al. (2014) provide a comprehensive survey of the many dimensions of the gap, from early-child differences to careers in academic science.

Studies of the gender gap have recently raised concerns about the gender neutrality of research evaluation promoted by public authorities and often used for public funding allocation to universities (Brooks et al., 2014). Large-scale research assessment is in place in many countries. The best known experience is the British Research Assessment Exercise (RAE), now revised and renamed Research Excellence Framework (REF), but research assessments with similar characteristics are now conducted in New Zealand, Australia, Hong Kong and Italy (Ancaiani et al., 2015), whereas in other countries universities funding is partly based on research performance indicators, for instance in Norway (Schneider, 2009), Denmark (Wright, 2014), and Czech Republic (Good et al., 2015).[1] Clearly, if the methodologies adopted in these research assessments are not gender neutral, they could provide negative feedback to academic institutions, with the unintended consequence of reinforcing the existing gender gap. Therefore, it is of paramount importance to verify whether this is actually the case, by analyzing results and methods used in these exercises.

* Corresponding author.
E-mail address: tullio.jappelli@unina.it (T. Jappelli).

http://dx.doi.org/10.1016/j.respol.2017.03.002
0048-7333/

[1] See http://www.arc.gov.au/era-reports for Australia, http://www.ugc.edu.hk/eng/ugc/rae/rae2014 for Hong Kong, and http://www.tec.govt.nz/Funding/Fund-finder/Performance-Based-Research-Fund-PBRF-/#Quality for New Zealand.

Moreover, research assessment provides an evaluation of the quality of a large sample of the research output of researchers of countries where they are conducted. Therefore, these exercises offer new data opportunities for studies on gender gap in research evaluation, especially if research output can be matched with researchers' characteristics and information on the evaluation process. In particular, evaluation exercises allow to investigate an important dimension of academic research, namely the judgment of peers on research quality, which is arguably one of the most important (and controversial) factor in determining career prospects.

In the British case, micro data are not available. Indeed, studies on RAE rely on department level information, matching aggregate RAE evaluations with department characteristics (Brooks et al., 2014; Taylor, 2011). In contrast, the Italian Agency for the Evaluation of Universities and Research (Anvur, the State Agency in charge of organizing the exercise) has direct access to micro level information on all researchers involved in the evaluation, research output as well as referees engaged in the peer review process. Moreover, the Italian evaluation of research (named VQR) provides an experimental environment to compare peer review and bibliometric evaluation, a key issue considering existing concerns about possible sources of discrimination against women in the evaluation process.[2]

Our analysis is based on this unique data set, and investigates the existence and magnitude of a gender gap in the quality of the research output, analyzing the evaluation results of all journal articles, monographs, book chapters and other research output ("research papers" for brevity) submitted to the evaluation. The dataset includes the best three research papers written between 2004 and 2010 by all Italian university professors and by all researchers employed by Italian public research institutes.

In the paper, we address four research questions. Our first hypothesis, given previous evidence on other dimensions of the gender gap, is that there is a gap also in research evaluation of published research. After measuring the gender gap, we discuss possible explanations for the measured gap, exploiting the rich amount of information available in the data set, and matching information on researchers and papers characteristics. The hypothesis is that the gap is lower or even disappears once we control for the professional rank of researchers. The reason is that the academic career and the quality of research output should be correlated, because scholars whose research is more appreciated by their peers should also be more likely to be associate or full professor, regardless of gender. The gap should also depend on number of co-authors and international collaborations because previous literature has shown that women have a disadvantage at networking as compared to men. One should also expect that the gap depends at least in part on gender differences in time and effort devoted to childcare. Moreover, we investigate if the evaluation method (bibliometric versus peer review) affects gender differences in measured research quality, and if the referees' gender affects the evaluation of women's research.

The Italian VQR is particularly suitable for this analysis in that it is a compulsory evaluation involving *all* staff with a permanent or temporary position in the universities and research institutes. Therefore, the analysis is not affected by self-selection or selection of researchers by the institutions involved in the assessment program. Moreover, the analysis applies to a large country with an homogeneous research environment: in Italy there is no distinction between research and teaching universities, and the hiring of academic staff is regulated at the national level, so that the average quality of the academic and research institutes is more homogeneous than in other countries (Abramo et al., 2012; Montanaro and Torrini, 2014). These features limit the scope for research staff selection and segregation according to the attitude towards research or teaching activities.[3]

The paper is organized as follows. Section 2 details the research questions and our contribution to literature. Section 3 describes the VQR exercise and the data used in the empirical analysis. Section 4 presents the main results, and reports evidence for the existence of a gender quality gap; it also investigate if the gap is related to observed characteristics of researchers, characteristics of research papers (i.e. journal article vs. books, book chapters, etc.), and other characteristics that could reveal a women disadvantage (i.e. the number of authors as a proxy for networking capacity). Section 5 further analyzes if the gap is affected by childcare and Section 6 if referee's gender affects the peer review process. Section 7 uses a random sample of papers evaluated by bibliometric analysis and peer review to detect the presence of gender bias in the evaluation method. Section 8 concludes.

2. Research questions

Glass ceiling is a major concern for academic career and a large body of research analyses the mechanisms that can explain the low presence of women in academic rankings, especially in the scientific fields. There is no general consensus on the relevance of the different factors at play. In particular, it is not clear if the gap can be explained by lack of a level playing field (for instance in terms of manuscript reviewing or grant funding) or by other factors that can affect women career choices or productivity.[4]

Many contributions have documented a gender productivity gap in research both in terms of number of published papers and citations.[5] Most of these studies, however, are restricted to specific scientific fields (Mauleón and Bordons, 2006) or research areas where research output is mainly in the form of papers published in English in indexed journals, where citation data are available (Abramo et al., 2009). Much less research covers areas such as humanities, law studies and social sciences where monographs and book chapters play an important role and where national language still prevails (Larivière et al., 2004; Larivière et al., 2006). Furthermore, research available in these fields tends to focus on journal articles, disregarding other research outputs (for instance West et al., 2013; Maliniak et al., 2013).

[2] HEFCE (2011), in a pilot exercise in preparation of the REF 2014 exercise, warns against the risk that the exclusive use of metrics to evaluate research could disadvantage women. The warning is based on a study that shows that papers authored by women are less cited. The pilot study does not investigate if fewer citations reflect the intrinsic quality of the papers and does not imply that women would attain higher evaluations with peer review. In contrast, Taylor (2011), based on results of the British RAE, expresses concerns about the exclusive use of peer review in that it could bias the evaluation in favor of some departments, when compared with bibliometric evaluations.

[3] In more heterogeneous university environments, women could be overrepresented in institutions with lower research intensity and research opportunities (Xie and Shauman, 1998).

[4] A recent comprehensive study on women in academic science (Ceci et al., 2014) concludes that gender discrimination is not a plausible explanation for the observed evidence and recent trends, and that more attention should be devoted to those factors affecting women choices before and after graduation. Larivière et al. (2013) express concern about the lack of a level playing field, calling for specific action aimed at improving the relative strength of women in research.

[5] See for instance, Larivière et al. (2013), Larivière et al. (2011) and West et al. (2013). Although the gap in terms of total citations is generally confirmed by a number of studies, the evidence on citations per work does not generally confirm the existence of a generalized gap (Ceci et al., 2014). Beaudry and Larivière (2016) argue that the citation rate may depend on productivity, so that in the fields where productivity is almost the same women show citation rates which are similar to those of men; however they would suffer from a negative gap in the fields where they are less productive.

Data on research assessment cover all scientific fields and all kind of research output. However, these exercises are usually limited to a subset of the research output of individual researchers in a given period (three papers in the VQR, four in the REF), and do not allow to obtain reliable measures of the productivity of researchers. Nonetheless, they allow to assess the existence and magnitude of a gender gap in terms of the quality of the best research produced in a given period, as measured by the judgment of peers through an explicit peer review process or indirectly in terms of citations and prestige of the journals where research is published.[6]

Therefore, they provide information on the way in which scientific communities judge the research of their members, both men and women, using criteria that are very similar to the ones used to assess the research quality of scholars for funding and promotion decisions. In this respect, they can detect a gender gap in a crucial dimension of the research profession, the judgement of peers, which is of paramount importance for the scientific career of scholars. Moreover, outcomes of research assessments are often used for the allocation of public resources, and are quite relevant for the governance of research institutions. It is therefore important that official research assessments provide a fair evaluation, because biased feedback to institutions could affect their hiring and promotion policies. Direct access to microdata of the Italian evaluation exercise allows us to investigate many dimensions through which women might be at a disadvantage in the course of an evaluation process.

The first research question of the paper is the analysis of gender differentials in terms of research quality, as assessed by the VQR. Although it is well established that women tend to be less productive in terms of research output than men, there is much less evidence on the relative quality of research performed by women. Research in this area has focused on the number of citations received by papers authored by women, sometimes as evidence of possible discrimination, sometimes as evidence of gender gap or advantage in research quality.[7] The quality of research, as evaluated by the peers, is arguably at least as important as its quantity, and a gap in this dimension would add (or reduce) the gap observed in terms of productivity.

Our first hypothesis is that, given a lower proportion of women among full and associate professors, possible disadvantages in research networking as documented or conjectured in a number of studies, and a possible gender bias in peers' evaluation both before and after publication, women will receive, on average, lower evaluations than men when comparing unconditional means or controlling only for demographics and field of study.

Moving forward to investigating the possible sources of the gender gap in research evaluation, we focus on researchers and publication characteristics. Our second hypothesis is on the role of personal characteristics, and in particular of the academic rank. To the extent that academic promotions reflect, at least in part, evaluation of peers in terms of research productivity and quality, we expect to observe a positive correlation between the rank of authors and the quality of their research. Considering that women are less represented among full professors, and to a lower extent among associate professors, we expect that any gap in terms of quality should decline or even disappear once we control for the rank of the author. Of course, this would not imply that women do not suffer from any disadvantage, but would point to factors that jointly determine their research quality (as evaluated by peers) and their career opportunities.

The third research question relates to the impact of some characteristics of published work, some of which are strictly linked to possible sources of disadvantage for women. As widely documented in the literature, women have more difficulties in networking and collaborations.[8] Therefore, we expect that the gap is partly explained by the number of co-authors and by the presence of international co-authors.

Our fourth hypothesis is that the gap depends at least in part on gender differences in time and effort devoted to childcare. In the labour economics literature on wage differentials between men and women this is one of the most investigated source of disadvantage. Indeed, government or contractual regulation on maternity leaves and availability of childcare facilities affect labour market outcomes of women, see for instance Arulampalam et al. (2007) and Del Boca (2015). As to the research gender gap literature, there is no consensus on the impact of family constraints, in particular childcare duties, on women performance in the academic profession. Fox (2005), Stack (2004), Krapf et al. (2014) find a negative impact on women productivity, whereas Joecks et al. (2014) find a positive relationship between fertility and academic output. It goes beyond the scope of this paper to address the complex interplay between family and academic career choice, considering the fundamental role that selection might play in this respect.[9] We limit our analysis to a comparison of the performance of women with children and those without. In a country like Italy where childcare support is notoriously weak and women on average still bear the most part of family duties (Del Boca, 2015, Anxo et al., 2011), we expect that the presence of children has a direct negative impact on the quality of research, as the presence of children may entail a reduction of the time spent in research.

To address these research questions our research strategy is straightforward. We compare women and men scores by regressing the score received by each research paper on a set of observable variables, and measure the gender gap through the estimated parameter on a dichotomous variable identifying the papers submitted by a woman. We exploit information on researchers'

[6] We do not enter into the debate of what research quality and how it can be measured, see Bonaccorsi (2015) for a comprehensive discussion. Here the concept of research quality coincides with the evaluation by peers, and we consider this judgement by itself quite relevant, as peers' judgement is the main driver of decisions regarding both individual scholars and institutions.

[7] We are not aware of comprehensive studies on gender differences based on national research assessments. Many studies compare the impact (number of citations) received by papers authored by women with papers authored by men. Maliniak et al. (2013) show that women's articles receive less citations in the field of International Relation, but the authors interpret this as a sign of discrimination rather than of a quality gap. On the contrary, Duch et al. (2012) show that in Ecology women receive more citations per paper than men (a sign of higher quality according to the authors) and do not detect any difference for chemistry. Beaudry and Larivière (2016) find that papers in health studies authored by women tend to receive fewer citations even controlling for papers characteristics and for the impact factor of the journal. Frandsen et al. (2015) find no evidence of a lower number of citations per paper in health studies. Ceci et al. (2014) in their survey conclude that there is no clear evidence that women receive, in general, less citations per paper than men.

[8] Brooks et al. (2014) find that co-authorship improves women performance in REF, and argue that difficulties in networking may hamper their ability to coauthor and their overall performance. Larivière et al. (2013) find that women tend to have less international collaborations than men. McNeely and Schintler (2010) analyze collaborations in STEMS, considering this issue as a key variable for improving women performance in science. McDowell et al. (2006) find gender differences in networking and publishing patterns in economics, but also report evidence of convergence with the increasing share of women in the field. On the other hand, Fell and König (2016) find no evidence of gender gap in collaborations among industrial-organizational psychologists.

[9] Ceci et al. (2014) argue that family and childcare duties mainly affect the choice to remain in the academic career. This means that most of the effect would be in the selection process rather than in the performance observed for those women who decide to stay in the academic pipeline. Similar conclusions are drawn by Ginther and Kahn (2006). Although it is beyond the scope of this paper to investigate the indirect impact of childbearing on women research performance through their career opportunities, we find evidence, that children in pre-schooling age reduce the probability of women to become professor (see Section 5).

characteristics (age, rank, gender, university affiliation, and scientific subject area), research output (type, number of authors, international co-authorship, and language) and method used in the evaluation (peer review vs. bibliometric analysis). The richness of the data set allows us to measure how much of the unconditional gender gap is explained by personal characteristics and how much is due to characteristics of research output.

In the second part of the paper, we investigate the role of the evaluation process itself. Many papers are concerned with the fairness of the evaluation methods in research as a possible source of discrimination and bias against women. In particular, some researchers explicitly advice against the use of metrics that could affect in a negative way the appraisal of women research (Brooks et al., 2014; HEFCE, 2011), or highlight a gender bias in citation patterns, see for instance Ward et al. (1992), Davenport and Snyder (1995), Larivière et al. (2013). Others show concern for the fairness of the peer reviewing process, although clear cut evidence of discrimination is quite difficult to be found apart from that based on experiments conducted in laboratory, as pointed out by Ceci et al. (2014).

To address this issue we exploit an experiment conducted during the VQR, where a sample of papers was evaluated by bibliometric indicators (based on a mixture of the journal's impact factor and citation analysis of individual papers) as well as peer review. Therefore we are able to compare, for the same set of publications, the gender gap observed when the score is based on bibliometric indicators with the one observed when this is based on peer review. Although the experiment does not allow us to draw any conclusion on the presence of discrimination in the citation patterns and in the peer review process, it allows us to investigate if any of the two methods appear to be more favourable to men or women than the other.

Finally, we analyse the role of a possible gender bias in the peer review process. Recent studies have tried to understand whether the performance of women can be related to the gender of evaluators, with mixed results. Some studies find that researchers benefit from the presence of same gender evaluators, see De Paola and Scoppa (2015). Others find an opposite gender preference among evaluators (Broder, 1993; Bagues et al., 2014), and still others find no significant role of gender (Zinovyeva and Bagues, 2011). We investigate this issue in detail by relating the score of papers evaluated by peer review to the gender of the referees.

3. Italian research evaluation and data

The Italian research evaluation exercise (or VQR) was carried out between the end of 2011 and July 2013 by the National Agency for the Evaluation of Universities and Research Institutions (ANVUR).[10] 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.[11] The purpose of the evaluation was to rank research institutions and departments within each research area based on the quality of their research.[12]

The evaluation was performed by 14 panels – one for each broad research area.[13] Each panel included an average of 32 researchers. Papers were evaluated using 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, according to the characteristics of the research field and of the research paper.[14] 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 evaluation was carried out by two external and independent reviewers, chosen by panel members taking account of conflicts of interest. In the case of diverging reports, the panel asked a third peer review or formed a consensus group to agree on a final score.[15]

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. Overall, in the VQR 53% of papers were evaluated by peer review and 47% by bibliometric analysis, with significant differences across areas. 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.[16] Peer review tends to prevail 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[17] and economics) was also evaluated by peer review. This implies that for a subset of about 7500 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.

Evaluation involves all the research staff formally affiliated to the institutions participating in the VQR.[18] This implies that VQR does not allow for staff selection for the submission of publications. It requires instead a selection of the best research papers. Academic researchers are required to submit their best three papers, while

[10] 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.

[11] For a detailed description of the VQR, see Ancaiani et al. (2015). For a comparison with the British REF see Rebora and Turri (2013) and Geuna and Piolatto (2016).

[12] Following other international experiences (Hicks, 2012), these rankings are currently used by the Ministry to allocate some 20% of the overall State funding to Italian public universities (Geuna and Piolatto, 2016; ANVUR, 2016).

[13] The 14 research areas are: (1) Mathematics and Computer Sciences; (2) Physics; (3) Chemistry; (4) Earth Sciences; (5) Biology; (6) Medical Sciences; (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.

[14] In scientific field, panels from 1 to 9 and Psychology, where journal articles prevail, evaluation was mainly based on bibliometric indicators. In the humanities, evaluation relied only on peer review. Books, book chapters and journal articles dealing with multidisciplinary or innovative issues on the border of different panels were evaluated by peer review.

[15] For further details regarding VQR, cfr. Ancaiani et al. (2015).

[16] 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 (e.g. papers published in 2010 for which available citations at the time of the research evaluation referred to only one year). We replicated the 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.

[17] Hard sciences correspond to Areas 1–9 (see fn. 13 for a definition of research areas).

[18] Unlike the British REF that allows institutions to select part of the research staff, i.e. the most productive and brilliant, for the submission of research papers, the Italian VQR requires all the research staff to participate in the evaluation. Each "missing" submission received a negative score of −0.5. Since the score of submitted papers ranged from 0 to 1, submitting a publication is strictly preferred to missing one. In the VQR the fraction of missing papers was about 5% for both men and women.

researchers employed by research centres submit six papers.[19] Therefore it is in the best interest of each department to guide each of its members in order to maximize the expected evaluation. This limits the risk of discrimination against women in the selection process of papers and does not allow the selection of staff. Only for papers co-authored by researchers belonging to the same institutions women could suffer from a gender selection bias. The VQR Call requires that papers with co-authors belonging to the same institution can be submitted only once, and that multiple submissions of co-authored papers are excluded from the evaluation. In case of co-authored publications by researchers affiliated to the same institution, department deans allocate the paper to only one researcher. In this allocation stage, discrimination may operate in the sense that departments may allocate the best papers to men instead of women, in case they have co-authored a paper. We investigate this possible source of discrimination in a robustness check where we focus on single-authored papers, for which the department has no role in the selection process.

Publications are evaluated with a qualitative score (excellent, good, acceptable, limited) which is then converted into a numerical scale ranging from 0 to 1. Papers classified "in the top 20% of the quality ranking shared by the international scientific community" are considered "excellent" and receive a score equal to 1, papers in the 60%-80% range are considered "good" and score 0.8, papers in the 50%–60% range are "acceptable" and score 0.5, while papers below the median receive "limited" and score zero. Each department's score is computed as the average score of all papers submitted by the department. For instance, the score of a department with 50 researchers is the average score computed over 150 papers (3 for each researcher).

Our sample includes data on almost 180,000 papers published in 2004–10 and submitted in early 2012 by all Italian universities and all public research centres to the VQR. 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 also on the gender of the two referees, their age and affiliation. 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. Men submitted 118,949 papers, or about two thirds of the sample, with the remaining third was authored and submitted by women.[20] 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 mature 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 age group the fraction of papers authored by men is 9 percentage points higher than for women). This pattern reflects a strong cohort effect, because access to academic positions by women has increased quite significantly over time. Indeed, women are far less well represented than men among full professors, who on average are older than associate and assistant professors.

Most papers submitted to the VQR are published in journals (74%), but there are also many 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 the sample, papers submitted by men have a higher probability of international co-authorship (23% vs. 19% for women). On average, 26% of the papers submitted by women are written in Italian, while for men the fraction is 21%. The proportion of single-authored papers is 25% for men against 31% for women.[21]

The proportion of papers submitted by men evaluated by bibliometric indicators was higher than for women, reflecting the higher proportion of journal articles submitted by men and the higher proportion of men in research areas where bibliometric analysis was more extensively adopted. Almost 81% of the papers were submitted by academic researchers, and this percentage is not different between men (80%) and women (82%). Furthermore, 26% of papers were submitted by institutions located in the South of Italy.[22] Finally, 74% of paper was submitted by large institutions (more than 600 researchers), 25% by medium size institutions and just 1% by small institutions.[23] These percentages do not differ by gender, hence we do not find a concentration of women or men by region or institution size.

Table 1 also reports the distribution by gender of the quality score. 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%. The difference in the average quality score between the two subsamples is 0.03 and is statistically significant at the 1% level, supporting the first hypothesis we lay down in Section 2. 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 7 percentage points higher for men (39% vs. 32% for women), and the difference is again statistically different from zero at the 1% level. 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 quality is almost entirely dependent on underrepresentation of women in the upper part of the quality score distribution.

4. The gender gap in research quality

In this section we check whether the unconditional difference in performance between men and women can be explained by the observable characteristics of papers and authors. In the first specification of Table 3 we report an ordered logit regression where the dependent variable is the quality score of each paper (taking 4 different values) and the independent variables are a dummy for gender and two age dummies. The regression also includes 15 research area dummies; standard errors are clustered at the level of the researcher submitting the paper.[24] We report results in terms of proportional odds ratio, because they have immediate and intuitive interpretation.

[19] The difference is explained by the implicit assumption that academic staff spends half of the working time in research and half in teaching activities.

[20] 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. The robustness of the results is checked controlling for the number of authors of each paper, and limiting the sample to single authored papers.

[21] Differences in publication practices partially reflect the higher incidence of women in humanities and social sciences. In the multivariate analysis in the next section we control for research field heterogeneity through a set of dummy variables indentifying the research area of the authors.

[22] Universities located in the South show, on average, a significant gap in terms of research quality (ANVUR, 2014) with respect to universities located in the Centre and in the North.

[23] Institution size is measured by the quartiles of the distribution of papers submitted to the VQR. Small institutions are those below the first quartile (166 papers, approximately 55 researchers); medium institutions are those between the first and third quartile (166–1900 papers); large institutions are those submitting more than 1900 papers (more than 630 researchers approximately).

[24] With respect to the 14 official research areas, the VQR splits Civil Engineering and Architecture into two separate areas. OLS regressions deliver similar results as ordered logit regressions.

Table 1
Sample statistics, by gender.

	Men		*Women*		*Total sample*	
	Mean	s.d.	Mean	s.d.	Mean	s.d.
Age less than 40 (dummy)	0.117	0.321	0.157	0.364	0.131	0.337
Age between 40 and 55 (dummy)	0.536	0.499	0.585	0.493	0.552	0.497
Age over 55 (dummy)	0.348	0.476	0.258	0.438	0.317	0.465
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)	0.334	0.472	0.518	0.500	0.397	0.489
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 types of research output	0.069	0.254	0.069	0.253	0.069	0.253
International co-authorship (dummy)	0.232	0.422	0.191	0.393	0.218	0.413
Research output in Italian language	0.214	0.410	0.260	0.439	0.230	0.421
Number of authors: 1 (dummy)	0.252	0.434	0.313	0.464	0.273	0.445
Number of authors: 2–5 (dummy)	0.458	0.498	0.401	0.490	0.439	0.496
Number of authors: 6+ (dummy)	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
University	0.799	0.401	0.817	0.387	0.805	0.396
South	0.251	0.434	0.262	0.439	0.255	0.436
Small size institution (less than 165 papers)	0.013	0.114	0.011	0.105	0.012	0.111
Medium size institution (between 165 and 1900 papers)	0.251	0.434	0.223	0.416	0.241	0.428
Numerical score	0.656	0.390	0.626	0.387	0.646	0.390
Number of observations	118949		61791		180740	

Table 2
Quality score, by gender.

	Men		*Women*		*Total sample*	
	No.	%	No.	%	No.	%
Excellent	46,457	39.1	19,701	31.9	66,158	36.6
Good	29,604	24.9	17,844	28.9	47,448	26.3
Acceptable	15,869	13.3	9409	15.2	25,278	14.0
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

In column 1 of Table 3 we find that for women, the odds of receiving a relatively high evaluation relative to low evaluations are 0.82 times lower than for men, holding constant the other variables in the model. Using the ordered logit estimates, we can also compute marginal effects and evaluate how the probability of receiving each evaluation changes with gender. Evaluated at the means of the other variables, we can see in column (1) of Table 4 that the probability of receiving the highest evaluation ("excellent") is 4.5 percentage points higher for men than for women; on the other hand, the probability of receiving the lowest evaluation ("limited") is 3.3 points higher for women.

The regression in column 1 of Table 3 also shows that younger researchers (less than 40 years old) have much higher odds of receiving good evaluations, relative to older researchers. Working in universities increases the odds of good evaluations by 64 percent, relative to researchers working in non academic institutions. For researchers working in the South and in small institutions the odds are considerably reduced with respect to researchers working in the North or in large institutions. These results support the first research hypothesis laid down in Section 2, even controlling for researchers' age and characteristics of research institutions. We also estimate different specifications adding as control variables the research sector (a finer classification with respect to the 15 research areas resulting in 353 sectors) and dummies for each of the 129 institutions. Results are similar to the specification reported in Table 3. In particular, the coefficient of the dummy for women is similar in magnitude and significance.

In the second specification of Table 3, we add as a control variable the rank of the researcher (associate professor and full professor, and equivalent positions in research institutes). The category of assistant professor is excluded. Being a full or associate professor increases considerably the odds of good evaluations. What is more interesting is that, holding constant the academic rank, the odds that women receive good evaluations relative to men increases to 0.95, considerably reducing the gender gap. This confirms our second research hypothesis, i.e. that the gap is lower once we control for the professional rank

In the third specification of Table 3 we check if the gender gap is explained by observable characteristics of research papers and by the evaluation method. In particular, we introduce dummies for publication type (book, book chapter, other research paper, while journal article is the excluded category), and a dummy for papers written in Italian. We control also for two variables that proxy for the ability or willingness to engage in networking activities: number of authors (2–5 and more than 5), and presence of an international co-author. Finally, we introduce a dummy for whether the paper was evaluated by peer review or bibliometric analysis.

Books, book chapters, and other research papers (e.g. designs, architectural plans, software, etc.) tend to receive lower evaluations relative to journal articles (the excluded category). Evaluations are higher for papers that are internationally co-authored, and increase with the number of co-authors. Papers written in Italian tend to receive lower scores than papers written in English or other languages. This is likely to reflect the fact that papers published in Italian journals are less widely disseminated than those published in English, and that in many research areas Italian journals publish less important results than international journals. Bibliometric evaluations are associated with a much larger odds ratio of receiv-

Table 3
The determinants of the quality score.

	Total sample			*Single-authored papers*
	(1)	(2)	(3)	(4)
Woman	0.822***	0.949***	0.938***	0.934***
	(0.0110)	(0.0129)	(0.0121)	(0.0202)
Age less than 40	1.943***	4.098***	3.032***	2.920***
	(0.0403)	(0.101)	(0.0710)	(0.117)
Age 40–55	1.515***	2.254***	1.937***	2.063***
	(0.0221)	(0.0361)	(0.0298)	(0.0531)
Full professor		3.104***	2.632***	3.535***
		(0.0597)	(0.0483)	(0.110)
Associate professor		1.647***	1.499***	1.643***
		(0.0265)	(0.0228)	(0.0433)
Book			0.953***	1.288***
			(0.0172)	(0.0322)
Book chapter			0.617***	0.820***
			(0.01000)	(0.0189)
Other research output			0.415***	0.796***
			(0.00838)	(0.0272)
International co-authorship			2.363***	
			(0.0372)	
Research output in Italian			0.434***	0.557***
			(0.00676)	(0.0124)
Number of authors: 2–5			1.195***	
			(0.0242)	
Number of authors: more than 5			1.748***	
			(0.0442)	
Bibliometric evaluation			5.089***	6.700***
			(0.0730)	(0.459)
University	1.640***	1.565***	1.443***	1.126
	(0.0362)	(0.0346)	(0.0301)	(0.173)
South	0.626***	0.619***	0.645***	0.641***
	(0.00884)	(0.00871)	(0.00876)	(0.0143)
Small size	0.779***	0.735***	0.746***	0.496***
	(0.0431)	(0.0416)	(0.0402)	(0.0422)
Medium size	1.105***	1.055***	1.050***	1.020
	(0.0160)	(0.0152)	(0.0144)	(0.0238)
Constant cut1	0.356***	0.830***	1.396***	0.973
	(0.0133)	(0.0328)	(0.0588)	(0.167)
Constant cut2	0.740***	1.766***	3.363***	3.085***
	(0.0276)	(0.0699)	(0.142)	(0.530)
Constant cut3	2.437***	6.040***	14.94***	25.88***
	(0.0913)	(0.241)	(0.638)	(4.460)
Observations	180,740	180,740	180,628	49,299

Note: The table reports odds ratios from ordered logit regressions for the quality score. Each regression includes 15 research areas dummies. Standard errors are clustered at the researcher's level. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

Table 4
The determinants of the quality score: marginal effects of gender.

		(1)	*(2)*	*(3)*	*(4)*
Excellent	Men	0.370	0.354	0.331	0.118
	Women	0.325	0.342	0.317	0.111
	Difference	0.045	0.012	0.014	0.007
Good	Men	0.289	0.298	0.356	0.411
	Women	0.288	0.298	0.356	0.401
	Difference	0.001	0.000	0.000	0.010
Acceptable	Men	0.142	0.148	0.154	0.251
	Women	0.154	0.151	0.159	0.256
	Difference	−0.011	−0.003	−0.005	−0.005
Limited	Men	0.199	0.201	0.159	0.219
	Women	0.232	0.209	0.168	0.231
	Difference	−0.033	−0.008	−0.009	−0.012

Note: The table reports marginal effects for men and women of the ordered logit regressions reported in Table 3, and the difference between the two. All margins are statistical significant at the 1% level.

ing high evaluations, relative to papers evaluated by peer review.[25] Most importantly, the odds that women receive higher scores is

[25] Only papers published in indexed journals are evaluated by bibliometric indicators (a combination of Impact Factor and citation counts). Therefore the higher odds ratio for bibliometric evaluation is due in part to a quality difference between two groups of papers. For instance, in scientific areas such as medical sciences, biology, chemistry, and physics, papers that are not published in indexed journals, are

essentially unaffected with respect to the regression of column 2 (marginal effects for each of the four quality scores are reported in Table 4). In short, the gender gap does not depend on the observable characteristics of papers submitted to the VQR.[26]

In the final specification of Table 3 we restrict the sample to single-authored papers. As mentioned in Section 3, one possible channel of discrimination is that departments might allocate the best co-authored papers to men instead of women. This seems not to be the case since the pattern of results in column 4 – and in particular, the gender effect – is quite similar to the full sample estimates. Overall, results in Table 3 don't support the third research hypothesis outlined in Section 2, that networking (as measured by co-authorship and international collaborations) affects the gender gap in research evaluation.

In Table 5 we report ordered logit regressions splitting the sample by academic rank (assistant, associate and full professor, or equivalent positions in research centres). The regressions are estimated using the most complete specification of Table 3. We find that the odds that women receive high evaluations relative to men is 0.94 for assistant professors, 0.95 for associate professors and 0.93 for full professors, holding constant the other variables in the model.

As a robustness check, in order to verify whether reviewers are more likely to 'defect' from the status of the journal for papers authored by women and whether the journal status cancel out any gender bias, we carry out further regression analysis. We consider as additional control variable bibliometric indicators that proxy journal quality and paper circulation in ISI-Web of Science and Scopus, see Table A.1 of the Internet Appendix. We use as indicator of journal status quartile dummies of the Impact Factor (IF) in the ISI database and of the Scimago Journal Ranking (SJR) in the Scopus database. As indicator of paper circulation and popularity, we use quartile dummies of total citation received in the VQR reference period (2004–2010) in the ISI and Scopus database. We conclude that a small gender bias in evaluation persists even controlling for bibliometric indicators. We also perform several robustness checks, for instance using the sample of journal articles or including in the sample also journals with no bibliometric indicators and an appropriate dummy. We find that the baseline results are unchanged.

We also replicate our analysis on a balanced sample obtained using a matching method (nearest neighbour) based on researchers' characteristics (gender, age, rank, geographical location of the institution and size) and find again the same results running propensity score weighted regressions only on common support observations (see Table A.2 of the Internet Appendix). Overall, the gender gap does not depend on the characteristics of the papers and of the evaluation method. However, the gap narrows substantially once we control for the academic rank of the author.[27]

In the next sections we explore other possible explanations and channels for the existence of a gender gap in the quality of research. We first check whether the gap is larger for women who have taken a maternity leave, and then explore if it could depend on gender discrimination by referees, or if it is affected by the evaluation method (bibliometric versus peer review evaluation).

5. Maternity leaves

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 or find a positive relationship between fertility and academic output.[28] Other studies identify motherhood choice and engagement in child care as prominent reasons for the underrepresentation of women in science (Ceci and Williams, 2011; Ginther and Kahn, 2006).

The literature provides little evidence of the effect of motherhood on research quality and evaluation, possibly due to lack of data. Brooks et al. (2014) find that the share of women that in the RAE exercise submitted less than four papers because of "individual circumstances" (in maternity leave, part-time worker, or early career researcher) does not affect the evaluation of British business schools. However, they find evidence that having "individual circumstances" affects women output rating as measured by the ranking of the journal they publish their papers, concluding that this could be driven by childbearing or part-time work. As explained in Section 2, in Italy childcare support is weak and women still bear the most part of family duties, so that we expect that the presence of children has a direct negative impact on the quality of research of women.

We can explore this question in the context of research quality evaluation by merging our dataset of papers and researchers with data on periods of leave provided by the Ministry of Education, Universities and Research (MIUR).[29] Data provision by universities and public research centres is voluntary and therefore may not include all researchers' leaves. The MIUR dataset includes data on about 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 increased 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. Based on the MIUR dataset, we can count the number of children a woman has and their age. In particular, we include in our specifications in Table 5 a dummy indicating whether a woman has at least one child of pre-schooling age in 2004–2010 (the reference period for the VQR). It should also be noted that the VQR attempts to compensate for maternity leaves by reducing by one the number of papers submitted for evaluation in case of maternity leaves up to 4 years, and by two in case of leaves over 4 years. Ex ante it is not clear whether this allowance compensates fully or partly for the leaves.

often less original and of lower impact, and were evaluated by peer review. The second and less important effect is that in our data bibliometric evaluation tends to be more generous than peer review. This point is highlighted in Cicero et al. (2013) who compare the two evaluations in the random sample of papers for which both evaluations are available that we will use in Section 7.

[26] Papers characteristics do not affect the gender gap even in a regression without controls for academic rank. Namely, controlling for papers characteristics in the regression in column (1) of Table 3 does not affect the gap.

[27] The validity of the regression analysis depends, among other things, on sample size. With a sample of about 180,000 observations and the characteristics that we have considered in the matching estimator, the problem of observing few women in some cells (for instance few women who are full professors in small university located in the South) should not be a major concern. Indeed, the average cell size is 1026 (only 23 cells have less than 20 observations).

[28] Fox (2005) finds that the productivity of women with preschool children is higher than that of women without children or those with school-aged children. Stack (2004) concludes that having children is not a strong predictor of productivity, and that the leading predictors of productivity are location in a research university and hours worked. Krapf et al. (2014) find that motherhood is not associated with low research productivity. Joecks et al. (2014) find that women in business and economics with children are more productive than women without children and that this is due to a selection effect that leads only most productive women to pursue an academic career and to have children at the same time.

[29] We thank MIUR for providing these data. The MIUR database contains the start and end date of each compulsory maternity leave. The merging of the data was anonymized.

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

	Full Professor (1)	*Associate Professor* (2)	*Assistant Professor* (3)
Woman	0.939**	0.954**	0.931***
	(0.0263)	(0.0218)	(0.0178)
Age less than 40	2.293***	3.381***	2.973***
	(0.385)	(0.175)	(0.107)
Age 40–55	1.732***	2.156***	1.970***
	(0.0441)	(0.0533)	(0.0642)
Book	0.910***	0.931**	1.032
	(0.0303)	(0.0298)	(0.0303)
Book chapter	0.632***	0.604***	0.612***
	(0.0187)	(0.0174)	(0.0163)
Other research output	0.451***	0.422***	0.381***
	(0.0177)	(0.0151)	(0.0120)
International co-authorship	2.440***	2.336***	2.341***
	(0.0766)	(0.0641)	(0.0571)
Research output in Italian	0.453***	0.437***	0.416***
	(0.0134)	(0.0120)	(0.0103)
Number of authors: 2–5	1.195***	1.202***	1.206***
	(0.0475)	(0.0419)	(0.0388)
Number of authors: more than 5	1.660***	1.814***	1.763***
	(0.0857)	(0.0794)	(0.0689)
Bibliometric evaluation	5.798***	5.172***	4.617***
	(0.166)	(0.133)	(0.0997)
University	1.600***	1.411***	1.435***
	(0.0731)	(0.0543)	(0.0430)
South	0.628***	0.640***	0.664***
	(0.0158)	(0.0154)	(0.0144)
Small size	0.673***	0.709***	0.820**
	(0.0703)	(0.0707)	(0.0646)
Medium size	1.017	1.050**	1.074***
	(0.0265)	(0.0255)	(0.0233)
Constant cut1	0.591***	0.968	1.327***
	(0.0441)	(0.0688)	(0.0897)
Constant cut2	1.440***	2.405***	3.124***
	(0.107)	(0.171)	(0.212)
Constant cut3	6.676***	10.96***	13.31***
	(0.500)	(0.789)	(0.914)
Observations	51,057	57,812	71,759

Note: The table reports odds ratios from ordered logit regressions for the quality score. Each regression includes 15 research areas dummies. Standard errors are clustered at the researcher's level. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

The first ordered logit regression in Table 6 indicates that the presence of children does not affect the odds of receiving a high score, other things equal, relative to men, women without children, or women with older children: contrary to our expectations, the odds of the variable "at least one child of pre-schooling age is, in fact, not statistically different from zero. One possible explanation for this result is that the VQR allowed women to submit only 2 papers instead of 3 (or even just one, in case of long or repeated leaves). Finally, regardless of the sign and significance of the maternity leave effect, controlling for leaves does not affect the coefficient of the dummy for women, and therefore the gender gap in research evaluation.[30]

The second regression restricts the sample to women, and the third regression to women less than 40 years old. Clearly in these regressions restricted to women we cannot measure any gender gap. However, it is of interest that the odds that women with children receive a relatively high score relatively to women without children, is 1.097 and statistically significant at 5% level, and 0.958 (but not statistically different from zero) in the sample of women under the age of 40. Although childbearing may affect the career opportunities and choices of women, it does not seem to be detrimental for their research quality and cannot explain the observed gender gap in VQR.

Although it is not the main focus of the paper, we also test whether having a child (or more than one) affects women's careers. This career effect could explain the result that childbearing has no effect on the evaluation of papers in the VQR, because selection operates well before the evaluation. That is, once a woman has chosen an academic career, it could be that having a child does not make a quality difference with respect to women without children. To this extent, we run a regression in which the dependent variable is a dummy for being a professor (associate or full) on a dummy for women, controlling for researchers' and papers' characteristics. Results available in Table A.3 of the Internet Appendix show that having at least one child of preschooling age is negatively associated with the probability of becoming professor (statistically significant-at the 1% level). This finding is consistent with previous literature (Ceci and Williams, 2011, Ginther and Kahn, 2006), and suggests that this mechanism is indeed at work also in our data.

6. Referee's gender

In this section we analyze how the presence of women among referees affects the evaluation of the quality of research as provided by the VQR. In particular, we explore whether men tend to discriminate women, and whether men or women tend to write more favourable evaluations of papers authored by other men or women.

[30] We tried different specifications using as alternative regressors the number of children of pre-schooling age or adding a dummy variable for the presence of children of schooling age. None of these coefficients is statistically different from zero.

Table 6
The effect of maternity leaves on research quality.

	Total sample (1)	*Women* (2)	*Women under 40* (3)
Woman	0.937***		
	(0.0124)		
Age less than 40	3.028***	2.521***	
	(0.0713)	(0.0960)	
Age 40–55	1.937***	1.807***	
	(0.0298)	(0.0480)	
Full professor (or equivalent)	2.633***	2.653***	3.113***
	(0.0483)	(0.0851)	(0.742)
Associate professor (or equivalent)	1.499***	1.503***	1.540***
	(0.0228)	(0.0362)	(0.111)
Book	0.953***	1.070**	1.237***
	(0.0173)	(0.0306)	(0.0753)
Book chapter	0.617***	0.670***	0.673***
	(0.01000)	(0.0168)	(0.0404)
Other research output	0.415***	0.506***	0.379***
	(0.00838)	(0.0170)	(0.0335)
International co-authorship	2.363***	2.389***	2.395***
	(0.0372)	(0.0649)	(0.139)
Research output in Italian	0.434***	0.489***	0.414***
	(0.00676)	(0.0118)	(0.0255)
Number of authors: 2–5	1.195***	1.150***	1.305***
	(0.0242)	(0.0388)	(0.101)
Number of authors: more than 5	1.747***	1.843***	1.906***
	(0.0442)	(0.0784)	(0.186)
Bibliometric evaluation	5.089***	4.761***	5.383***
	(0.0730)	(0.118)	(0.299)
At least one child in pre-schooling age	1.017	1.097**	0.958
	(0.0395)	(0.0443)	(0.0539)
University	1.442***	1.381***	1.631***
	(0.0302)	(0.0510)	(0.133)
South	0.645***	0.624***	0.611***
	(0.00876)	(0.0140)	(0.0296)
Small size	0.746***	0.625***	0.794
	(0.0402)	(0.0576)	(0.123)
Medium size	1.050***	1.029	1.068
	(0.0144)	(0.0240)	(0.0534)
Constant cut1	1.396***	1.472***	0.515***
	(0.0588)	(0.108)	(0.0754)
Constant cut2	3.361***	3.617***	1.342**
	(0.142)	(0.266)	(0.195)
Constant cut3	14.93***	17.81***	6.497***
	(0.638)	(1.328)	(0.953)
Observations	180,628	61,760	11,810

Note: The table reports odds ratios from ordered logit regressions for the quality score. Each regression includes 15 research areas dummies. Standard errors are clustered at the researcher's level. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

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 of discrimination or 'gender preference'. Some studies find that researchers benefit from the presence of same gender evaluators (De Paola and Scoppa, 2015). Others find an opposite gender preference among evaluators (Broder, 1993; Bagues et al., 2014), and yet others find no significant role of gender (Zinovyeva and Bagues, 2011).

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. In the VQR, peer review evaluation is organized as follows. Panel members assign each paper to two external referees chosen independently by two experts of the panel. The referee report was organized around three questions (originality, relevance, and international outreach), and each provided a scored ranging from 1 to 9, which were then aggregated into a single score (ranging from 3 to 27). The scores of the two referees were averaged to obtain a single score, and finally converted by the panel into a final merit class (limited, acceptable, good, excellent), on the same scale as the bibliometric evaluation. The peer review was performed on published papers, and was therefore "informed" and single-blinded: the evaluators were fully aware of the identity and hence of the gender of the authors. It is precisely this feature of the VQR that allows us to verify whether referees have biased attitude towards women. A second characteristic of the dataset is that the number of referees is quite substantial (almost 15,000 in the different research areas).

The first column of Table 7 reports the proportional odds ratio from the estimation of an ordered logistic regression where the dependent variable is the averaged quality score of the two referees' evaluations. The sample includes only peer reviewed papers (97,414 observations). The specification uses the same variables as in our baseline regressions, adding as control variables the referees' characteristics: the sum of the ages of the two referees, whether both referees are affiliated to an Italian institution, and dummies for the gender of the two referees (whether the paper was evaluated by one or by two women). The regression includes also two interaction terms between the gender of the researcher submitting the paper, and the dummies for the gender of the referees.

The results suggest that, on average, women tend to give more generous evaluations than men. The odds ratios of the dummy variables "one referee is a woman" and "two referees are women" are indeed larger than one (1.12 and 1.168, respectively) and statisti-

Table 7
The effect of referees' gender and of the evaluation method.

	Sample of all peer reviewed papers	*Random sample*	
		Peer review score	*Bibliometric score*
Woman	0.968	0.802***	0.939
	(0.0192)	(0.0474)	(0.0515)
Age less than 40	2.874***	2.871***	3.852***
	(0.0827)	(0.257)	(0.389)
Age 40–55	1.963***	1.882***	2.176***
	(0.0358)	(0.112)	(0.145)
Full Professor (or equivalent)	2.902***	2.080***	2.820***
	(0.0643)	(0.147)	(0.223)
Associate Professor (or equivalent)	1.528***	1.263***	1.501***
	(0.0280)	(0.0715)	(0.0917)
Book	0.935***		
	(0.0195)		
Book chapter	0.557***		
	(0.0104)		
Other research output	0.383***		
	(0.00881)		
International co-authorship	2.429***	2.254***	2.292***
	(0.0601)	(0.116)	(0.129)
Research output in Italian	0.447***	0.0935***	0.172***
	(0.00795)	(0.0258)	(0.0348)
Number of authors: 2–5	1.111***	1.137	1.334**
	(0.0264)	(0.140)	(0.160)
Number of authors: more than 5	1.661***	1.480***	1.634***
	(0.0542)	(0.195)	(0.215)
Sum of age the two referees	1.002***	0.999	
	(0.000443)	(0.00162)	
Both referees are affiliated to an Italian institution	0.764***	0.943	
	(0.0118)	(0.0467)	
Both referees are women	1.168***	1.161	
	(0.0396)	(0.208)	
One referee is a woman	1.120***	1.137**	
	(0.0202)	(0.0733)	
Woman × both referees are women	1.025	1.243	
	(0.0465)	(0.312)	
Woman × one referee is a woman	0.966	1.081	
	(0.0271)	(0.116)	
University	1.415***	1.258***	1.703***
	(0.0418)	(0.0823)	(0.121)
South	0.660***	0.610***	0.616***
	(0.0108)	(0.0345)	(0.0376)
Small size	0.597***	1.351	1.175
	(0.0400)	(0.325)	(0.353)
Medium size	1.025	0.942	0.993
	(0.0173)	(0.0534)	(0.0626)
Constant cut1	1.118	0.447***	0.982
	(0.0839)	(0.111)	(0.172)
Constant cut2	3.373***	1.223	1.573***
	(0.253)	(0.302)	(0.275)
Constant cut3	22.82***	11.81***	4.518***
	(1.738)	(2.944)	(0.793)
Observations	97,414	7407	7453

Note: The table reports odds ratios from ordered logit regressions for the quality score. Regression (1) is estimated on the sub-sample of papers evaluated by peer review. Regressions (2) and (3) are estimated on the random sample evaluated by both peer review and bibliometric analysis. Each regression includes 15 research areas dummies. Standard errors are clustered at the researcher's level. Standard errors are reported in parentheses. (***), (**), (*) denote statistical significance at the 1%, 5%, and 10% level, respectively.

cally different from zero at the 1 percent level. However, the more generous evaluations of women are not directed towards a specific gender. Indeed, the odds ratios of the coefficients of the interaction terms "*Woman* × *Both referees are women*" and "*Woman* × *One referee is a woman*" are not statistically different from zero. It is also of interest that the odds ratio that women receive high evaluations relative to men (0.968) is not much affected in this sample of peer reviewed papers relative to the full sample estimates (although the estimate has a larger standard error). This shows that the gender gap holds also in a restricted sample that excludes articles published in indexed journals evaluated in the VQR on the basis of bibliometric indicators.[31]

As a robustness check, we run a regression at the level of each referee, therefore considering each paper as two separate observations, using as dependent variable the referee's score (ranging from

[31] In order to take into account that evaluation by a woman is not random but depends on the scientific field and the relative presence of women, we use a Heckman estimator. In the first stage we explain the probability that the referee is a woman as a function of the proportion of women in the different scientific areas and the characteristics of the paper. In the Heckman model the correlation between the unobservable variables and the selection coefficient are not statistically different from zero. In comparison with the baseline model, the results are qualitatively unaffected, as shown in Table A.4 of the Internet Appendix.

3 to 27), introducing referees' fixed effects, and clustering errors at the paper level. In these regressions the coefficient of the variable "*Woman* × *Referee is a woman*" is small, but positive and statistically different from zero. The difference between the two types of regressions is that in the order logit reported in Table 7 we measure the impact on the probability of having higher evaluations of a four classes scale, whereas in the OLS estimates we use the evaluation of the referees on a 3–27 scale. We conclude that in the logit regressions the impact of the interaction terms are irrelevant to jump from a lower to a better evaluation in the four merit classes of the VQR, see Table A.5 of the Internet Appendix.

7. Bibliometric evaluation vs. peer review

Two further channels of discrimination are worth exploring. First of all, since 74% of the referees were men, it might be that women are discriminated by the peer review process. Second, to the extent that women are less cited than men, and that in many areas the majority of scientists are men, bibliometric evaluation, which is largely based on citation analysis, might be lower for women. Obviously, any judgement that discriminates explicitly against women would break the rules of the evaluation, which clearly requires fairness, but this is precisely what a regression analysis is able to detect.

Some papers find that women receive less citations per article published than men when controlling for the characteristics of the paper and of researchers, see for instance Davenport and Snyder (1995), HEFCE (2011) and Beaudry and Larivière (2016). Larivière et al. (2013) find 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. According to Brooks et al. (2014) the use of journal metrics to score papers could penalize women. Using a very complete database of funding, scientific papers and citations compiled at the individual researchers' level in Quebec, Beaudry and Larivière (2016) find that when women collaborate with the same number of co-authors as men, or target similar Impact Factor journals, their articles are less cited then those of their male colleagues.

Studies of gender discrimination are based on citation data and do not compare bibliometric evaluation with peer review. This comparison is interesting, because it allows to check if peer review evaluations attenuate (or worsen) the disadvantage that women may face with bibliometric evaluation. We are able to make this comparison exploiting a distinctive and quite useful feature of the VQR. For statistical comparison, a random sample of 10% of all papers evaluated by bibliometric analysis was evaluated also by peer review.[32] This sample of nearly 7500 papers was stratified by research areas, and includes all areas in which bibliometric indicators were used in the evaluation (it therefore excludes humanities, law, and sociology). The random sample allows a statistical comparison between the two evaluation methods, and in particular, the degree of agreement between bibliometric evaluation and peer review.[33]

The last two regressions of Table 7 report the ordered logit regressions for the quality score using as dependent variable the peer review score (column 2) and the bibliometric score (column 3), based on a combination of journal impact factor and citations received by individual papers.[34]

Both regressions signal the presence of a gender gap. In this sample the odds that women receive high evaluations with bibliometric indicators is 0.94 relative to men, but not statistically different from zero. Instead, the odds that women receive high evaluations with peer review, relative to men, is much lower (0.8), and statistically different from zero at the 1 percent level. In the regression for peer evaluation we control for the gender of the referee, and therefore the result cannot be explained by discrimination against women by men. Overall our findings do not support concerns about the possible negative impact of the use of bibliometric indicators on women's evaluation in the VQR. Although access to high impact factor journals and citation patterns may suffer from gender bias, we do not find evidence that peer review evaluation would attenuate the problem. Rather, there is evidence that bibliometric evaluations tend, on average, to be more favourable to women than peer review. Our results are obtained using a particular selection of academic publications i.e. the three best papers selected by the author published in 2004–10. Therefore, they do not extend automatically to the entire scientific production of researchers.

8. Conclusions

This paper contributes to the literature on the gender gap in research evaluation. We exploit a large dataset of roughly 180,000 papers evaluated during the 2004–2010 Italian evaluation of universities and research institutions. The dataset provides detailed information on the type of publication, the evaluation method (peer review or bibliometric analysis), and the characteristics of authors and referees. Moreover it contains the results of the evaluation (a quality score), and how this evaluation was performed (either by referee reports and peer review evaluation, or relying on citation analysis).

The empirical analysis suggests the existence of a significant gender gap in research evaluation, and supports the hypothesis that the gap is reduced if one controls for the academic rank of researchers. Instead, there is not supporting evidence that networking effects damage the evaluation of women research, and that family responsibilities played a role in the outcomes of the Italian research evaluation. In particular, in our baseline estimate, we find that for women the odds of receiving a relatively high evaluation versus relative low evaluations are 0.82 times lower than for men. Controlling for university rank (assistant, associate, full professor or equivalent), gender inequality falls sharply but the gap persists and cannot be explained by research output characteristics (type of publication, number of authors, international collaborations, language of publication).

In the rest of the paper we check several variables that might explain the gap, focusing on parental leaves for childbearing, to understand whether the gap is larger for women who experienced maternity leave, or if it stems from discrimination. The results suggest that maternity leaves do not play a major role in explaining the gap, although they affect the career opportunities of women and indirectly their research performance.

Identifying the presence of discrimination is difficult since it can take many forms. We explore two potential sources: discrimination against women by referees, and discrimination against women by bibliometric evaluation. We analyze carefully the sub-sample of peer reviewed papers and find some that women provide more generous evaluations, but the coefficient of the interaction between

[32] The final evaluation of these papers was based on the bibliometric indicators, and peer review reports were collected only for statistical purposes.

[33] Cicero et al. (2013) report detailed statistics by research area on the difference between the two types of evaluations. Bertocchi et al. (2015) compare different evaluation methods in economics, management, and statistics.

[34] The VQR bibliometric evaluation criteria assigned to each journal article indexed in ISI-WoS or Scopus a score linked to the percentiles of the journal citation impact and of the citations received.

"referee is a woman" and "paper authored by a woman" is not statistically different from zero. Thus, a woman does not improve, as a referee, women's evaluations more than the evaluations of papers authored by men.

A random sample of journal articles evaluated by both peer review and bibliometric indicators provides further insights. A statistical comparison between the two evaluation methods reveals that bibliometric evaluations are comparatively more generous for women than peer review evaluation: the gender gap is higher if the paper is evaluated by peer review. This result should attenuate concerns on the use of bibliometric indicators as a possible source of gender gap in research evaluations like the Italian VQR or the British REF. In fact, we find that peer review evaluation is less favourable to women than evaluations based on a combination of citations and indicators of journal ranking.

Overall, we find no evidence that the VQR has been "unfair" to women. But it is important to stress that the focus of this study is on a particular gap, arising in the evaluation of research by peers at a given stage of researchers' careers. It might well be that gender differences emerge at other stages of the career, and even well before entry in academic or research institutions. The finding of a persistent (albeit small) gap might be due to a "selection effect" rather than an "evaluation effect". The variable that is more strongly correlated with the size of the gap is the academic ranking, since once we control for the rank the residual gap is considerably reduced. This means that higher attention should be devoted to understanding the career path of women, since their undergraduate studies or even before, along the lines suggested by Ceci et al. (2014), trying to identify the factors that affect research quality, as perceived by peers, as well as their career opportunity. Furthermore, our results are obtained using a selected sample of papers (the best three papers published in 2004–10), and in future research it would be useful to use a more extensive sample of academic publications.

The empirical analysis suggests that the largest contributor of the gap is the low probability of obtaining top evaluations. This is consistent with an explanation of the residual gap based on women behaviour, as suggested by the literature on gender differences in preferences, according to which women seem to be more risk averse than men (Borghans et al., 2009), tend to engage less in competitive behaviour (Niederle and Vesterlund, 2007; De Paola et al., 2015), and display a lower propensity to specialise their research (Leahey, 2006), an attitude that might limit their ability to achieve a high reputation among peers and to publish in the most prestigious journals.[35]

The findings provide a broad overview of the research quality evaluation of women relative to men, and show that some of the concerns regarding the evaluation of women research might have been overstated by the current literature, and that more effort should be devoted to understand women attitudes and career obstacles that could affect women choices and results since the early stages of their career.

In future research it would be interesting to explore in more detail our rich dataset to highlight possible heterogeneities across research fields. In particular, it would be interesting to study how the observed gap is affected by the density of women and research practices in the different fields.[36] Indeed, there is substantial variability in the degree of internationalisation (for instance humanities and law studies tend to be more inward oriented than hard sciences), publication practices (article journals vs. monographs and book chapters), collaborations (in hard sciences research is organized in teams, much less in the humanities), funding (hard sciences require more complex and expensive infrastructures). In principle, each of these characteristics might affect women and men in different ways.

It would be also interesting to explore in more detail why women obtain more variable results than men, and have a lower probability to obtain an excellent evaluation, and in particular if this could be related to a less focused strategy with respect to men. Finally, it would be interesting to study how the gender gap changes over time and if the adoption of the VQR, which creates a clear incentive for institutions to improve their research performance, is contributing to reduce or increase the gap. On this front, data from the new VQR will be soon avaliable, and will allow estimates of gender gap in research evaluation also in 2011–14, and comparison between different time periods.

Acknowledgments

We thank the Editor and three anonymous referees for many helpful insights and comments, and Anvur for providing the data.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.respol.2017.03.002.

References

Anvur, 2014. Rapporto sullo stato del Sistema universitario e della ricerca 2013. http://www.anvur.it.
Anvur, 2016. Rapporto sullo stato del Sistema universitario e della ricerca 2016. http://www.anvur.it.
Abramo, G., D'Angelo, C.A., Caprasecca, A., 2009. The contribution of star scientists to overall sex differences in research productivity. Scientometrics 81 (1), 137–156.
Abramo, G., Cicero, T., D'Angelo, C.A., 2012. The dispersion of research performance within and between universities as a potential indicator of the competitive intensity in higher education systems. J. Informetr. 6 (2), 155–168.
Ancaiani, A., Anfossi, A.F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., Cicero, T., Ciolfi, A., Costa, F., Colizza, G., Costantini, M., Di Cristina, F., Ferrara, A., La Catena, R.M., Malgarini, M., Mazzotta, I., Nappi, C.A., Romagnosi, S., Sileoni, S., 2015. Evaluating scientific research in Italy: the 2004–10 research evaluation exercise. Res. Eval. 24 (3), 242–255.
Anxo, D., Mencarini, L., Pailhé, A., Solaz, A., Tanturri, M.L., Flood, L., 2011. Gender differences in time-use over the life-course in France, Italy, Sweden, and the U.S. Feminist Economics 17 (3), 159–195.
Arulampalam, W., Booth, A.L., Bryan, M.L., 2007. Is there a glass ceiling over Europe? Exploring gender pay gap across the wage distribution. Ind. Labor Relat. Rev. 60 (2), 163–186.
Bagues, M., Sylos-Labini, M., Zinovyeva, N., 2014. Do gender quotas pass the test? Evidence from academic evaluations in Italy, Scuola Superiore Sant'Anna, LEM Working Paper Series, 14.
Beaudry, C., Larivière, V., 2016. Which gender gap? Factors affecting researchers' scientific impact in science and medicine. Res. Policy, in press.
Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C.A., Peracchi, F., 2015. Bibliometric evaluation vs. informed peer review: evidence from Italy. Res. Policy 44 (2), 451–466.
Bonaccorsi, A., 2015. La valutazione possibile. Teoria e pratica nel mondo della ricerca. Il Mulino, Bologna.
Borghans, L., Heckman, J.J., Golsteyn, B.H., Meijers, H., 2009. Gender differences in risk aversion and ambiguity aversion. J. Eur. Econ. Assoc. 7 (2–3), 649–658.

[35] To check if women are more risk averse than men in their choice of research projects, we collapse the data by researcher, and regress the variance of the quality score against a dummy for women and other researchers' characteristics. The coefficient of the dummy for women is positive and statistically different from zero, pointing to a larger variability of research quality of women. This result is more consistent with a less focused research strategy, and a stronger propensity to undertake multidisciplinary projects, rather than with higher risk aversion by women. Results are available in Table A.6 of the Internet Appendix. It should also be noticed that these results should be taken with great care, because the number of observations on which we can compute the standard deviation of individual scores is limited to three observations for most of the sample. We thank an anonymous referee for suggesting this regression.

[36] Although the sample size is large (about 180,000 observations), in some disciplines the number of women is rather small, so that there are only few observations for some categories.

Broder, I.E., 1993. Review of NSF economics proposals: gender and institutional patterns. Am. Econ. Rev. 83 (4), 964–970.
Brooks, C., Fenton, E.M., Walker, J.T., 2014. Gender and the evaluation of research. Res. Policy 43 (6), 990–1001.
Ceci, S.J., Williams, W.M., 2011. Understanding current causes of women's underrepresentation in science. Proc. Natl. Acad. Sci. 108 (8), 3157–3162.
Ceci, S.J., Stephen, J., Ginther, D.K., Kahn, S., Williams, W.M., 2014. Women in academic science: a changing landscape. Psychol. Sci. Public Interest 15 (3), 75–141.
Cicero, T., Malgarini, M., Nappi, C.A., Peracchi, F., 2013. Bibliometric and peer review methods for research evaluation: a methodological appraisement, MPRA Paper, (50470).
Davenport, E., Snyder, H., 1995. Who cites women? Whom do women cite? An exploration of gender and scholarly citation in sociology. J. Doc. 51 (4), 404–410.
De Paola, M., Scoppa, V., 2015. Gender discrimination and evaluators' gender: evidence from Italian academia. Economica 82 (325), 162–188.
De Paola, M., Ponzo, M., Scoppa, V., 2015. Gender differences in attitudes towards competition: Evidence from the Italian scientific qualification, CSEF Working Papers, (391), http://www.csef.it/WP/wp391.pdf.
Del Boca, D., 2015. Child care arrangements and labor supply, IDB WORKING PAPER SERIES No. IDB-WP-569.
Duch, J., Zeng, X.H.T., Sales-Pardo, M., Radicchi, T., Otis, S., Woodruf, T.K., Amaral, L.A.N., 2012. The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. PLoS One 7 (12), e51332.
European Commission, 2015. She Figures 2015. https://ec.europa.eu/research/swafs/pdf/pub_gender_equality/she_figures_2015-final.pdf.
Fell, C.B., König, C.J., 2016. Is there a gender difference in scientific collaboration? A scientometric examination of co-authorships among industrial-organizational psychologists. Scientometrics 108 (1), 113–141.
Fox, M., 2005. Gender family characteristics, and publication productivity among scientists. Soc. Stud. Sci. 35 (1), 131–150.
Frandsen, T.F., Jacobsen, R.H., Wallin, J.A., Brixen, K., Ousager, J., 2015. Gender differences in scientific performance: a bibliometric matching analysis of Danish health sciences graduates. J. Infometr. 9, 1007–1017.
Geuna, A., Piolatto, M., 2016. Research assessment in the UK and Italy: costly and difficult but probably worth it (at least for a while). Res. Policy 45 (1), 260–271.
Ginther, D.K., Kahn, S., 2006. Does science promote women? Evidence from academia 1073–2001, NBER Working paper, No. 12691.
Good, B., Vermeulen, N., Tiefenthaler, B., Arnold, E., 2015. Counting quality? the czech performance-based research funding system. Res. Eval. 24 (2), 91–105.
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 Issues paper, February 2011/03.
Hicks, D., 2012. Performance-based university research funding systems. Res. Policy 41 (2), 251–261.
Joecks, J., Pull, K., Backes-Gellner, U., 2014. Childbearing and (female) research productivity: a personnel economics perspective on the leaky pipeline. J. Bus. Econ. 84 (4), 517–530.
Krapf, M., Ursprung, H.W., Zimmermann, C., 2014. Parenthood and productivity of highly skilled labor: Evidence from the groves of academe, IZA Discussion Papers No. 7904.
Larivière, V., Lebel, J., Lemelin, P., 2004. Collaborative research in the social sciences and humanities: Bibliometric Analysis of Practices, Report to the Social Sciences and Humanities Research Council of Canada (SSHRC), Observatoire des sciences et des technologies, Resource document.
Larivière, V., Archambault, E., Gingras, Y., Vignola-Gagné, È., 2006. The place of serials in referencing practices: comparing natural sciences and engineering with social sciences and humanities. J. Am. Soc. Inf. Sci. Technol. 57 (8), 997–1004.
Larivière, V., Vignola-Gagné, E., Villeneuve, C., Gélinas, P., Gingras, Y., 2011. Sex differences in research funding productivity and impact: an analysis of Québec university professors. Scientometrics 87 (3), 483–498.
Larivière, V., Ni, C., Gingras, Y., Cronin, B., Sugimoto, C.R., 2013. Global gender disparities in science. Nature 504 (7479), 211–213.
Leahey, E., 2006. Gender differences in productivity research specialization as a missing link. Gend. Soc. 20 (6), 754–780.
Maliniak, D., Powers, R., Walter, B.F., 2013. The gender citation gap in international relations. Int. Organ. 67 (04), 889–922.
Mauleón, M., Bordons, M., 2006. Productivity impact and publication habits by gender in the area of Materials Science. Scientometrics 66 (1), 199–218.
McDowell, J.M., Singell, L.D., Stater, M., 2006. Two to tango? Gender differences in the decisions to publish and co-author. Econ. Inq. 44, 153–168.
McNeely, C.L., Schintler, L., 2010. Gender issues in scientific collaboration and workforce development: implications for a federal policy research agenda. In: Workshop on the Science of Science Measurement, U.S. Office of Science and Technology Policy, Washington, DC, Retrieved from http://www.nsf.gov/sbe/sosp/workforce/mcneely.pdf.
Montanaro, T., Torrini, R., 2014. Il sistema della ricerca pubblica in Italia, Bank of Italy Occasional papers n. 219.
Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? do men compete too much? Q. J. Econ. 122, 1067–1101.
OECD, 2015. Education at a Glance 2015. http://www.oecd-ilibrary.org/education/education-at-a-glance-2015_eag-2015-en.
Rebora, G., Turri, M., 2013. The UK and Italian research assessment exercises face to face. Res. Policy 42 (9), 1657–1666.
Schneider, J.W., 2009. An outline of the bibliometric indicator used for performance-based funding of research institutions in Norway. Eur. Polit. Sci. 8 (3), 364–378.
Stack, S., 2004. Gender children and research productivity. Res. High. Educ. 45 (8), 891–920.
Taylor, J., 2011. The assessment of research quality in UK universities: peer review or metrics? Br. J. Manage. 22 (2), 202–217.
Ward, K.B., Gast, J., Grant, L., 1992. Visibility and dissemination of women's and men's sociological scholarship. Soc. Probl. 39 (3), 291–298.
West, J.D., Jacquet, J., King, M.M., Correll, S.J., Bergstrom, C.T., 2013. The role of gender in scholarly authorship. PLoS One 8 (7), e66212.
Wright, V.S., 2014. Knowledge that counts: points systems and the governance of Danish universities. In: Alison, I., Griffith D, Smith, E. (Eds.), Under New Public Management: Institutional Ethnographies of Changing Front-line Work. University of Toronto Press, pp. 294–338.
Xie, Y., Shauman, K.A., 1998. Sex differences in research productivity: new evidence about an old puzzle. Am. Sociol. Rev. 63, 847–870.
Zinovyeva, N., Bagues, M., 2011. Does gender matter for academic promotion? Evidence from a randomized natural experiment, IZA Discussion Paper No.5537.