Oswald, Andrew J.

Working Paper

Can we test for bias in scientific peer-review?

IZA Discussion Papers, No. 3665

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Oswald, Andrew J. (2008) : Can we test for bias in scientific peer-review?, IZA Discussion Papers, No. 3665, Institute for the Study of Labor (IZA), Bonn, https://nbn-resolving.de/urn:nbn:de:101:1-2008090165

This Version is available at:
https://hdl.handle.net/10419/35220

DISCUSSION PAPER SERIES

IZA DP No. 3665

Can We Test for Bias in Scientific Peer-Review?

Andrew J. Oswald

August 2008

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Can We Test for Bias in Scientific Peer-Review?

Andrew J. Oswald
University of Warwick,
Cornell University and IZA

Discussion Paper No. 3665
August 2008

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3665
August 2008

ABSTRACT

Can We Test for Bias in Scientific Peer-Review?*

Science rests upon the reliability of peer review. This paper suggests a way to test for bias. It is able to avoid the fallacy – one seen in the popular press and the research literature – that to measure discrimination it is sufficient to study averages within two populations. The paper's contribution is primarily methodological, but I apply it, as an illustration, to data from the field of economics. No scientific bias or favoritism is found (although the Journal of Political Economy discriminates against its own Chicago authors). The test's methodology is applicable in most scholarly disciplines.

JEL Classification: H8

Keywords: discrimination, citations, science, peer-review system

Corresponding author:

Andrew J. Oswald
Department of Economics
University of Warwick
Coventry CV4 7AL
United Kingdom
E-mail: andrew.oswald@warwick.ac.uk

* For excellent research assistance, I thank Joao Tovar Jalles. For their helpful comments, I thank Danny Blanchflower, Nick Bloom, Bill Dickens, Glenn Ellison, Hamning Fang, Amanda Goodall, Dan Hamermesh, Kirabo Jackson, David Laband, Larry Katz, Peter Neary, Charles Oppenheim, Scott Smart, Justin Wolfers, and Stephen Wu. Thanks also go to Cornell University for its hospitality, and to the ESRC for research support.

Can we Test for Bias in Scientific Peer-Review?

Andrew J Oswald
University of Warwick and Cornell University

1. Introduction

Is the peer-review system biased? This question matters both for science and public policy. Most of modern knowledge is predicated upon the reliability and intellectual objectivity of that system.

In this paper I try to suggest a test for discrimination by scientific journals. It builds on the intuition that in a discriminatory world those from a minority who make it through a system have to be better than the norm. At the heart of the test is the use of accumulated citation totals on a set of paired adjacent articles. Citations[1] are taken as a proxy for the objective quality of an article (measured with the benefit of hindsight), and I focus on adjacent articles on the assumption that editors order their articles approximately by perceived quality, with the highest at the top of a journal issue. Each of the two articles in an adjacent pair has jostled for the same ordered space in the issue, and thus, by the nature of the selection process, is at that stage viewed by the editor as of equivalent quality ex ante. It seems likely that many articles cannot be ranked unambiguously by an editor; but a weak ordering[2] is all that is required for the later test.

In this way, the paper argues, it is possible to explore the idea that editors systematically under-estimate -- relative to an article's true quality -- the scientific contribution of articles being submitted by a particular kind of author or from a particular part of the world. The citations data provided by history go on to reveal

[1] It thus falls in a tradition represented by work such as Hamermesh et al (1982), Oppenheim (1995, 2007), Hamermesh and Schmidt (2003), and Goodall (2006), in which citations are treated as important, real signals. This paper does not claim, however, that citations are free of error, nor that in the long run it will be sensible to see citations as unambiguously valuable (the more that citations data are emphasized, the more probable it is that their signalling value will gradually be eroded by opportunistic behaviour).

[2] It does not matter for the test if editors choose the order in a random way; the only problem would be if they deliberately chose to put better papers lower down the issue. In revising this paper, I was told of one journal where the policy is to allocate all but the top slot in an issue on the basis of articles'

which articles outperform their close neighbors within an issue of a journal. As those data accrue, they allow us to learn whether the editor's judgment[3] was systematically biased. I thus build upon the idea that "an editor, while uncertain about the future impacts of the papers that are submitted, should … be publishing articles whose expected impacts are identical" (Hamermesh, 2002).

Because the paper's test operates within-journal, *it is robust to the common concern that certain journals are over-cited* because of their fame. Moreover, the test *does not require* that the supply curve of articles from a (favored) majority group be identical to that from the minority[4].

The objective of this paper is to suggest a testing procedure that other researchers, in any scholarly discipline, might find straightforward to use. But, as an illustration, I implement one application to two of the world's currently most-cited economics journals, the Quarterly Journal of Economics and the Journal of Political Economy[5]. This later part of the paper uses historical data on 302 QJE articles to check for the existence of international bias and pro-Harvard bias, and data on 208 JPE articles to check for the existence of pro-Chicago bias.

Mention should be made of an important unpublished paper by Smart and Waldfogel (1996). This was initially unknown to me but was drawn to my attention after early drafts of the current paper. The underlying idea proposed by Smart and Waldfogel is identical to that studied later. Although there are differences in the implementation method[6], and the authors do not use an adjacency test *per se*, I would like to

---

chronological order of submission. Such a rule also satisfies the requirements needed later for the suggested test.

[3] I use 'editor' to stand, as shorthand, for the combination of editorial decision-making and the advice being given by referees.

[4] I emphasize these two because they were often raised by readers of early drafts.

[5] Wu (2007) discusses the possibility of bias. He shows that just 4 universities account for approximately 40% of the recent papers published in the QJE, and that this kind of concentration has risen through the decades.

[6] Rather than, as later, the use of a chi-squared test on adjacent articles, the authors estimate regression equations in which citations are the dependent variable and the independent variables include the order number of the article. These two approaches should give similar results if (i) the regression equation includes issue dummies, (ii) the lead articles are omitted from the sample, and (iii) the length of journal is held constant. Under some further assumptions, restriction (ii) will also be unnecessary.

emphasize that the Smart-Waldfogel paper greatly predates my own. Later in the paper I discuss their substantive findings.

2. Background

Throughout social science, and especially labor economics[7], there is a literature that uncovers examples of discrimination in the world (a particularly clear demonstration is Goldin and Rouse 2000, who show that female musicians are rated more highly, controlling for quality, if heard from behind an anonymizing screen). Despite the avowed disinterestedness of decision-making in the university world, some researchers feel that the journal system is similarly unreliable[8], and is biased against minorities.

Link (1998) finds, in her study of approximately 4000 submissions to the journal Gastroenterology, that US referees exhibit a marked preference for papers written by US authors rather than for those by non-American authors. Einav and Yariv (2006) and Van Praag and Van Praag (2008) document a form of surname bias. Budden et al (2007) show that after the journal Behavioral Ecology went over to double-blind refereeing the proportion of female-authored papers accepted by the journal rose strongly. Equivalent concerns are heard across a number of academic settings (documented in psychology, for example, in Blackburn and Hakel 2006, and in management studies by Macdonald and Kam 2007). Some European scientific researchers suggest -- see for example the discussion in Luwel 1999 -- that the major US journals discriminate against them.[9]

---

[7] Cain (1986) surveys the early literature. An interesting recent example of a discrimination test is by Wolfers (2006) who studies -- but finds no evidence in support of -- the hypothesis that female-headed companies produce systematically better results than the stock market expects.

[8] Although now fractionally dated, Amstrong (1997) is an impressively careful summary of empirical evidence on the quality of peer review.

[9] Hudson (2007) argues that citation levels are partly due to chance events such as which other article is in the issue of the journal. Frey (2003) and Starbuck (2005) are doubtful of quality control within elite journals, and Tsang and Frey (2006) question the increasingly prescriptive nature of the refereeing system. However, Laband et al 2002 provide evidence that quality control in the subject of economics is reasonably good. Ellison (2007) questions whether traditional peer review will be undermined by the internet. Neary et al (2003) and Oswald (2007) also discuss the quality of modern economics research.

The paper attempts to contribute to this literature by suggesting an empirical method that does not require detailed knowledge of the individual accept-and-reject decisions made inside editorial offices[10]. Arguably one advantage of the paper's pairing test is the unusually mild data requirements that it makes. It is unnecessary to measure the citations to, or other characteristics of, the majority of articles in a journal issue. Moreover, the test's focus on contiguous articles within an issue means that the influence of changes in the nature of the journal and editorial style, and of alterations in research fashion through the years, are helpfully minimized.

3. The Averaging Fallacy

A common error -- it might be termed the averaging fallacy -- in informal debate on discrimination is to focus on the averages in two populations, such as a favored group and a minority, and to argue that because the mean value of variable-of-interest X is higher in one group than the other then this is evidence of injustice or inefficiency or both. In a large class of cases such an approach is conceptually wrong. The reason is that fairness and efficiency will typically require that it is the marginal values of variable X that should be equated. Averages may not be informative[11] about those marginal values.

Consider the following analytical example where for concreteness there is a choice to be made by an editor about how many 'home' versus 'foreign' articles to publish. Publishing more of one type then inevitably means rejecting more of the other.

Assume that a journal receives submissions from two sources, a large and familiar home group of researchers and a smaller foreign group of researchers. Quality of articles is q, and lies by definition between zero and unity, where unity is the best scientific work that is feasible. The quality distribution of home articles is h(q) and that of foreign articles is f(q). The length of the journal, namely the number of

[10] It is likely that editors would, perhaps reasonably, reject requests to allow this kind of scrutiny in their offices; referees are promised anonymity when they take on the task of acting as reviewers. Hence another route has to be found, using only the revealed choices made by the journal.

[11] This is sometimes known in the literature as the infra-marginality problem. In their discussions of possible racial discrimination against students in school and college admissions, Dickens and Kane (1999) and Bowen and Bok (2000) discuss this kind of error. See also, in a different setting, Anwar and Fang (2006).

articles that can be accepted for publication, is fixed at K. Assume that editors or referees may act in a discriminatory way by unfairly weighting foreign work less highly, ceteris paribus, than home work. Let the degree of bias be captured by a coefficient b. In an unbiased world, therefore, b is zero. The parameter b can be viewed as the downgrading percentage adjustment that is implicitly or explicitly made in evaluating the work of the unfamiliar minority. Whether knowingly or unknowingly, a journal editor takes b as given[12], but otherwise acts to maximize the total quality of the articles published in the journal. The editor chooses cut-off quality level alpha on the home papers and beta on the minority papers to solve

$$Maximize: E = \int_{\alpha}^{1} qh(q)dq + \int_{\beta}^{1} q(1-b)\,f(q)dq \qquad (1)$$

$$s.t.: \int_{\alpha}^{1} h(q)dq + \int_{\beta}^{1} f(q)dq = K. \qquad (2)$$

At a maximum, this leads, with a non-negative multiplier λ, to first-order conditions for the two quality thresholds:

$$\alpha: -\alpha + \lambda = 0 \qquad (3)$$

$$\beta: -\beta(1-b) + \lambda = 0 \qquad (4)$$

and therefore to the kind of marginal condition typical in economic analysis, namely, that for the observed outcomes among articles published in the journal,

$$\alpha = \beta(1-b) \leq \beta \qquad (5)$$

Quality required of home authors $\leq$ *Quality required of foreigners*

[12] It seems probable that editors believe b is zero in their particular journal. Clark and Wright (2007) contains an interesting summary of editors' views on fairness.

or, in other words, that the minimal acceptable quality of an article from a foreign author is equal to an effective beta that has been corrected upwards for a 'tax' of b. Hence alpha is (weakly) below the quality required of the foreign authors.

Equation (5) is a formal statement of the familiar idea that in order to compete a minority group has, in general, to be better than those individuals who are in the majority.

Because it has not been emphasized in the bibliometric literature as much as is desirable, it seems useful to stress one point. It is not possible to test for the existence of discrimination empirically by calculating the mean values of quality in each group. Knowing the difference in group means, given by,

$$D = [\int_{\alpha}^{1} qh(q)dq / \int_{\alpha}^{1} h(q)dq] - [\int_{\beta}^{1} qf(q)dq / \int_{\beta}^{1} f(q)dq] \quad (6)$$

is in general unrevealing about whether condition (5) holds, because D can be positive or negative while still being consistent with (5), and in the kind of case described above there is no reason to think that the distributions h(.) and f(.) will be identical[13].

To go beyond impasse, some way has thus to be found to explore the marginal condition (5) and to assess empirically the size of the discrimination coefficient, b. To do this, the paper draws on the fact that journal editors have to order their choice of articles, after they have been accepted for publication. Historically this was because journals appeared only in print form, and editors were thereby required to decide, as a matter potentially of importance to authors and readers, which article should go first in an issue, which second in that issue, and so on. Human beings notice things at the top of a page. Common sense and human nature suggest that editors had, and in the electronic age presumably still have, a tendency to put at the head of the ordering those articles that they view -- necessarily ex ante -- as those most likely to be important. It is known empirically, for example, that lead-papers in

[13] By definition, the minority group are few in number, and when compared to the more homogenous majority are likely to be an unpredictable mixture of high talent and low talent, and of highly resourced and poorly resourced.

journal issues tend to attract more citations (Smart and Waldfogel 1996; Hudson 2007), and this may be because they are intrinsically better. Coupe et al (2008), however, find -- using a natural experiment -- that the citations effect is small. Judge et al (2007) contacted 16 editors of journals and found that the great majority said they did attempt to put at the top of the issue the article they viewed as best.

The revealed choices by editors provide extra information for those who wish to investigate possible scientific bias. Assume that in ordering the articles in his or her journal an editor puts the best work systematically at the start. He or she is assumed, in a journal issue with 10 slots, to adopt the following rule: place an article of perceived quality q10 at higher or equal to that in the within-issue ordering compared to an article of perceived quality q9, and above or equal to one of q8, and so on. A weak ordering suffices for the later test. Given this assumption, two conclusions can then be drawn about adjacent articles:

(i) the prior assessed quality of article qi is greater than or equal to the assessed quality of article qi-1

(ii) except where the number of articles is small, and the market thus 'thin', the assessed quality of article qi will be similar to the assessed quality of article qi-1.

Given this foundation, a version of equation (5) can be implemented. By the revealed choices of the journal, articles next to each other can be taken to have the characteristics described by (i) and (ii). This is because the adjacent articles have successfully competed for approximately the same ordered space in the issue, and thus, by the nature of the review and editorial process, can be viewed as of approximately equal ex ante quality.

We are then interested in whether article qi-j has more or fewer citations than the contiguous article qi-j-1. The natural test statistic is thus:

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

which is the familiar chi-squared test where e is expected frequencies and o is observed ones.

4. An Application to the Field of Economics

As an illustration, I study the possibility of two kinds of bias[14]. I first collect data on the Quarterly Journal of Economics from 1970 to 2002 (ending the data series there in order to allow some years for citations to accrue). The journal is published by Harvard University Press and accepts few papers from non-Americans. It is an important journal. At the time of writing, the QJE has the highest impact factor[15] in the subject of economics. The data were collected in one week in May of 2008. I searched, for example, on the word England in the Web of Knowledge, found 109 articles, and proceeded from there to inspect each of these, and the articles around them. The aim is to test initially the idea that the QJE is biased against a particular set of economists, those from English universities. The test is then applied to check for favoritism towards Harvard authors, and for discrimination against researchers in a set of European countries. Articles were discarded according to the following rules. First, for the England and European tests, all articles co-authored with others such as North Americans (more precisely, anyone with a US address) were eliminated. This was to allow an inquiry into the effect of purely non-American authors submitting their work. Second, Comments, Replies and Notes were eliminated. This was to avoid the difficulties of comparisons with full-length articles. Third, articles were excised if they were the first in the issue, although in practice this was especially rare for English economists' papers. The reason for the rule in this case is because, for the international discrimination test, a comparison is required with the article that comes immediately before the article of interest.

The data and results are given in the Appendix[16]. The author's name is stated first; then the year; then the lifetime citations total; then data on whether the article was more highly cited than the one immediately prior to it in the issue, and then a 'yes' or 'no'. In the first test, for England-based authors, the answer 'yes' means that that article is more highly cited than the previous, adjacent article. This is a sign that,

[14] Hamermesh (2002) studies the impact in a citations regression equation of a variable for the country of origin of the author.

[15] It is approximately the same as that for the Journal of Economic Literature, but I leave that publication aside because it does not publish original scientific papers of a conventional sort.

[16] The huge variation in the degree of cited-ness in even this elite journal is clear, and this phenomenon has been noted before by authors such as Starbuck (2005) and Oswald (2007).

despite the editor's choice of order, the England-based research garnered more subsequent attention.

The principal observations for the anti-England test (76 usable observations)

The result, after the eliminations from the data set of 109 observations, was
38 pairs; total YES = 12; total NO=26
and thus for a set of 38 pairs between 1970 and 2002 authored solely by people with an English address. These were examined alongside the article listed immediately before it in the journal. The citations total of the English-authored article was compared to that of this prior, adjacent article. A chi-squared test was done.

The findings from this pairing test turn out to be reassuring for journal editors and those concerned about the integrity of the peer review system. Of the total articles in the QJE sample, 26 were preceded by a more highly cited article, and 12 were preceded by a less highly cited article (there were no ties). This is different from 50:50 but the null hypothesis of that division cannot quite be rejected at conventional confidence levels. More importantly, insofar as there is any bias, it works in the opposite direction from that expected. Perhaps most tellingly, the direction of this division is what would be predicted for an efficient system where the earlier articles were if anything marginally better than -- within each pair -- the second of the adjacent articles. Hence there is no evidence, using the paper's suggested test, for discrimination by the peer-review system of the Quarterly Journal of Economics against research articles[17] emanating from England.

The principal observations for the pro-Harvard test (172 usable observations)

I next do a test for pro-Harvard bias. For sharpness, I concentrate on solely-Harvard authors. Here the data come out as
86 pairs; total YES = 38; total NO = 48

[17] This is despite the fact, to which a naive discrimination theorist might point, that only 38 of nearly 2000 QJE articles were (solely) from England.

This means, technically, that the data are consistent with a slight amount of bias in favor of Harvard authors. But the extent of the division is only marginally different from 50:50 and the numbers do not allow the null hypothesis of no discrimination to be rejected. The chi-squared test value with one degree of freedom is here 50/43, which is greater than the 0.455 critical number that would be generated randomly 50% of the time, but far below the 2.706 number required for significance at the 10% level, and approximately equivalent to significance at the 30% level. Hence the Harvard-discrimination finding is not statistically significant at any conventional confidence level.

The principal observations for the anti-European test (54 usable observations)

This takes a group of European nations, other than England, given by the set =
[France+Spain+Italy+Germany+Switzerland+Belgium+Sweden+Holland]
And the test result is: *27 pairs; total YES = 12; total NO =15.*

This implies that there is no evidence of international, anti-European discrimination. Once again, any bias, in this small sample, goes in the 'wrong' direction.

A final check is to turn to a different journal. I attempt an equivalent test for the Journal of Political Economy, which is produced by the University of Chicago Press. In this case the test is symmetric to the pro-Harvard test, but is now a Chicago-favoritism test.

The principal observations for the pro-Chicago test (208 usable observations)

Here the data come out as
104 pairs; total YES = 72; total NO = 32

This means that the data are consistent with a large amount of bias, but this is, paradoxically, *against* Chicago authors. The chi-squared test value with one degree of freedom is approximately 8, which implies that the null is rejected at the 0.005 significance level. Hence these data suggest there is no discrimination by the JPE in

favor of Chicago authors. The reverse is the case: the JPE apparently sets a higher standard for its own home authors.

These results are consistent with those derived in a different way, necessarily on much older data, in the creative paper by Smart and Waldfogel (1996). The authors do not find evidence for either international discrimination or for gender discrimination; they uncover some support for the idea that editors treat articles by authors from low-ranked institutions more favorably than those by authors at top-20 schools. Nor do they find evidence of favoritism at the QJE, although, unlike this paper's result, they conclude that Harvard authors are held to a higher standard than outsiders.

5. Discussion

The implementation of the test has the nature of a one-zero form. Although it might be argued that mean citations of the two sets of articles could also be compared, doing this makes no difference (as would be predicted by the theoretical view that the contiguous articles are of similar quality) to the paper's main empirical conclusion. There seems to be no evidence, for example, that the QJE discriminates[18] against English or European authors, or in favor of the home Harvard-located authors. This finding is reminiscent of one in innovative research by Hamermesh (2002), who establishes, by examining major journals, that empirical papers by non North Americans are not cited disproportionately heavily. However, his paper examines mean values, and in principle is open to the objection that D in equation (6) is not a reliable guide to the existence of discriminatory behaviour at the margin.

A final consideration is whether we have set up the null hypothesis inappropriately as that of no discrimination (that hypothesis has not been rejected in the English, Harvard or European cases). Might it be that, especially for the Harvard test, we are making a Type II error and the null is actually false?

---

[18] This result now seems to me natural, but I had expected the reverse finding. The reasons may perhaps be instructive -- in illustrating how views of discrimination are shaped by randomness, internal self-justification, and small numbers of observations. My (three) articles accepted by the QJE each came when I had an American affiliation, and in retrospect I am conscious of the papers I submitted that were both rejected by that journal and went on to be rather influential, but have somewhat forgotten those articles I submitted that were both rejected and went on to be rather un-influential. Such difficulties of perception are natural, but they may be widespread.

To see why this is particularly unlikely in the English case, consider the possibility that the true model is that the ratio of outcomes should be 26 No to 12 Yes (in the answers in the data table). What we actually observe is 12 Yes and 26 No. To get a feel for the likelihood, imagine moving wrongly from 12:12 to 12:26. The probability of this is 0.5 to the power 14, which is approximately 0.0001. It is only when we approach a true model where the degree of journal bias against foreigners is positive but very small, that we begin to reach a point where the English data would find it difficult to reject this alternative null of bias. In the Harvard example, of course, it is not possible to reject the null of mild favoritism. But it is not clear why that is the appropriate null. In the Chicago case, there is a form of negative favoritism.

Another, and arguably rather natural, way to think about the QJE data is as each of the three approaches being a single test that should be pooled. Then, the three applications of the paper, when combined into one, produce the overall finding: *the combined data set of 302 QJE observations reveals, where randomness would predict a division of all our pairs of 75.5:75.5, an observed split of 72:79*. This is almost an even division in the data (and in fact tilts in the wrong direction to be consistent with bias), and thus why the chi-squared tests fail and why there is no evidence of bias. The Yes and No of the Harvard test are reversed in meaning from the Yes and No of the other two for England and Europe. The asymmetry is because we are then testing for Harvard favoritism and for international bias (so one is a form of discrimination pro-something, and the other is a form of discrimination anti-something).

6. Conclusion

This paper proposes a way to test for bias in scientific peer-review. The test can be applied in any setting where hierarchically-ordered choices are made; it works by examining the lifetime citations to pairs of contiguous journal articles[19]. It is able to avoid the fallacy -- one often seen in the popular press and sometimes in the large

[19] Like Smart and Waldfogel (1996), which is a precursor to this paper, the test is one for the existence of discriminatory behaviour by journals. It is not a test of discrimination by a whole society (there then being no objective scientific criterion such as lifetime citations), even though such a test would be interesting if it could be constructed.

discrimination research literature -- that to measure discrimination it is sufficient to examine averages within two populations.[20]

The paper's method draws on information from the chosen ordering of articles in a journal issue, and uses that to draw inferences about editors' unmeasured beliefs about quality. Although the main contribution of the paper is methodological, the procedure is implemented, as an illustration, to test for two kinds of discrimination in data on 510 articles from 1970-2002 on two leading journals of economics, the Quarterly Journal of Economics and the Journal of Political Economy. Virtually no evidence for discrimination by the QJE is found. Put more precisely, no evidence is uncovered for international bias against authors from English or European universities. A tiny amount of evidence consistent with pro-Harvard bias is found. But this is not a statistically significant effect on a chi-squared test. If the three QJE sub-tests within the paper are thought of as being elements of a single one, the combined 302 observations produce an observed division for the 151 pairs of 72:79, which is almost identical to the 75.5:75.5 split that randomness would imply. This is why the chi-squared tests do not reject the null of no discrimination[21].

For the JPE, the paper's adjacency test uncovers the reverse of what might have been expected. The data reveal that this Chicago journal acts in a way that discriminates against its own. For home authors, the bar is apparently set higher.

The paper's test can be applied to other forms of discrimination (such as on grounds of race or gender or style of research). In principle, it can be used by investigators

[20] To my knowledge, there have been only a few attempts to confront the difficulties caused in discrimination testing by this infra-marginal problem (Dickens and Kane 1999 being one).

[21] It is natural, then, to ask why many researchers believe that journals act in discriminatory ways. A possible explanation is that journals make frequent mistakes. Evidence for the randomness of journals is provided in Starbuck (2005) and Oswald (2007); that unreliability is visible in the variation, even within an issue, in the citations numbers to the QJE and JPE articles studied here. It seems that referees are bad at forecasting which papers will go on to be important. Hence we can all look back on papers incorrectly rejected by referees too 'biased' to recognize the iconoclastic nature of our contributions. For authors -- perhaps especially for young authors -- who try to make sense of their rejection letters, these mistakes by journals may be interpreted as something dark and calculatingly systematic, while the truth will often be the more mundane and embarrassingly human one of random error.

with minimal[22] statistical knowledge. This form of adjacency test appears to have potential applications in a range of disciplines and academic settings.

[22] It could be used, for example, by a humanities scholar who knows how to look up a chi-squared table.

References

Anwar, Shamena. and Fang, Hanming. 2006. An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. American Economic Review, 96, 128-151.

Armstrong, J Scott. 1997. Peer review for journals: Evidence on quality control, fairness, and innovation. Science and Engineering Ethics, 3, 63-84.

Blackburn, Jessica L. and Hakel, Milton D. 2006. An examination of sources of peer-review bias. Psychological Science, 17, 378-382.

Bowen, William G. and Bok, Derek. 2000. The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions. Princeton University Press, Princeton.

Budden, Amber E., Tregenza Tom, Aarssen, Lonnie W., Koricheva, Julia, Leimu Roosa, and Lortie, Christopher J. 2007. Double-blind review leads to increased representation of female authors. Trends in Ecology & Evolution, 23, 4-6.

Cain, Glen C. 1986. The economic analysis of labor market discrimination: A survey. Handbook of Labor Economics, Elsevier, Amsterdam.

Clark, Timothy C. and Wright, Michael. 2007. Reviewing journal rankings and revisiting peer reviews: Editorial perspectives. Journal of Management Studies, 44 (4), 612-621.

Coupe, Tom, Ginsburgh, Victor, and Noury, Abdul. 2008. Are leading papers of better quality? Evidence from a natural experiment. Working paper. Kyiv School of Economics.

Dickens, William T. and Kane, Thomas J. 1999. Racial test score differences as evidence of reverse discrimination: Less than meets the eye. Industrial Relations, 38 (3), 331-363.

Einav, Liran and Yariv, Leeat. 2006. What's in a surname? The effects of surname initials on academic success. Journal of Economic Perspectives, 20 (1), 175-188.

Ellison, Glenn. 2007. Is peer review in decline? Working paper, MIT. July.

Frey, Bruno S. 2003. Publishing as prostitution? Choosing between one's own ideas and academic success. Public Choice, 116, 205-223.

Goldin, Claudia and Rouse, Cecilia. 2000. Orchestrating impartiality: The impact of "blind" auditions on female musicians. American Economic Review, 90 (4), 715-741.

Goodall, Amanda H. 2006. Should research universities be led by top researchers, and are they? A citations analysis. Journal of Documentation, 62 (3), 388-411.

Hamermesh, Daniel S. 2002. International labor economics – Presidential address. Journal of Labor Economics, 20 (4), 709-732.

Hamermesh, Daniel S., Johnson, George E. and Weisbrod, Burton A. 1982. Scholarship, citations and salaries: Economic rewards in economics. Southern Economic Journal, 49(2), 472-481.

Hamermesh, Daniel S. and Schmidt, Peter. 2003. The determinants of Econometric Society fellows elections. Econometrica, 71, 399-407.

Hudson, John. 2007. Be known by the company you keep: Citations - quality or chance? Scientometrics 71(2), 231-238.

Judge, Timothy A., Cable, Daniel M., Colbert Amy E., and Rynes, Sara L. 2007. What causes a management article to be cited? Article, author, or journal? Academy of Management Journal, 50, 491-506.

Laband, David N., Tollison, Robert D. and Karahan, Gokhan Ramazan. 2002. Quality control in economics. Kyklos, 55, 315-334.

Link, Ann M. 1998. US and non-US submissions. Journal of the American Medical Association, 280(3), 246-247.

Luwel, Marc. 1999. Is the science citation index US-biased? Scientometrics, 46(3), 549-562.

Macdonald, Stuart and Kam, Jacqueline. 2007. Ring a ring o' roses: Quality journals and gamesmanship in management studies. Journal of Management Studies, 44, 640-655.

Neary, J. Peter, Mirrlees, James A. and Tirole, Jean. 2003. Evaluating economics research in Europe: An introduction. Journal of the European Economic Assocation. 2003, 1, 1239-1249.

Oppenheim, Charles 1995. The correlation between citation counts and the 1992 Research Assessment Exercise Ratings for British library and information science university departments. Journal of Documentation, 51, 18-27.

Oppenheim, Charles 2007. Using the h-index to rank influential British researchers in information science and librarianship. Journal of the American Society for Information Science and Technology, 58, 297-301.

Oswald, Andrew J. 2007. An examination of the reliability of prestigious scholarly journals: Evidence and implications for decision-makers. Economica, 74, 21-31.

Smart, Scott and Waldfogel, Joel. 1996. A citation-based test for discrimination at economics and finance journals. Working paper, Indiana University, and NBER paper 5460. January.

Starbuck, William H. 2005. How much better are the most prestigious journals? The statistics of academic publication. Organization Science, 16, 180-200.

Tsang, Eric W. K. and Frey, Bruno S. 2006. The as-is journal review process: Let authors own their ideas. Working paper, University of Zurich.

Van Praag, Mirjam. and Van Praag, Bernard M.S. 2008. The benefits of being economics professor A (and not Z). Economica, forthcoming.

Wolfers, Justin. 2006. Diagnosing discrimination: Stock returns and CEO gender. Journal of the European Economic Association, 4, 531-541.

Wu, Stephen. 2007. Recent publishing trends at the AER, JPE, and QJE. Applied Economics Letters, 14(1), 59-63.

APPENDIX

The principal observations for the QJE England test (where Yes indicates discrimination)

TOTAL = 109 results; after exclusions = 38 results; total YES = 12; total NO=26

Author(s)	Year	Total Cites	#Cites of BEFORE	Higher cites than before (Y/N)
DEVLETOG.NE	1971	6	3	YES
ROBINSON S	1971	44	8	YES
ATKINSON AB	1973	21	695	**NO**
ROBINSON	1975	24	74	**NO**
RAU N	1975	1	1	**NO**
EATWELL J	1975	8	41	**NO**
LECOMBER R	1977	1	98	**NO**
HART OD	1977	8	1	YES
PISSARIDES CA	1978	18	84	**NO**
NGUYEN DT	1979	3	4	**NO**
WATERSON M	1980	15	169	**NO**
AKERLOF GA	1980	201	59	YES
LORIE HR	1980	0	4	**NO**
SEN A	1981	14	53	**NO**
HART O	1982	169	17	YES
NORMAN G	1983	12	52	**NO**
VENABLES AJ	1983	3	6	**NO**
KEHOE TJ	1985	24	0	YES
GRUBB D	1986	2	28	**NO**
DEMEZA D, WEBB DC	1987	97	5	YES
KLEMPERER P	1987	170	170	**NO**
MEYER MA	1987	6	12	**NO**
NAYLOR R	1989	33	65	**NO**
LAYARD R, NICKELL S	1990	31	20	YES
FRANK J	1990	1	6	**NO**
PISSARIDES CA	1992	55	7	YES
TIMMERMANN AG	1993	40	44	**NO**
ANDERLINI L, FELLI L	1994	25	50	**NO**
MEYER MA	1994	9	123	**NO**
MANNING A	1995	18	66	**NO**
VANREENEN J	1996	30	35	**NO**
BURDETT K, COLES MG	1997	61	33	YES
BATEMAN I, MUNRO A, RHODES B, STARMER C, SUGDEN R	1997	61	287	**NO**
DE MEZA D, LOCKWOOD B	1998	44	29	YES
BLUNDELL R, PRESTON I	1998	42	66	**NO**
BESLEY T, BURGESS R	2000	17	49	**NO**
CAROLI E, VAN REENEN J	2001	44	27	YES
VIOLANTE GL	2002	19	42	**NO**

The principal observations for the QJE Harvard test (where No indicates discrimination)

TOTAL = 208 results; after exclusions = 86 results; total YES = 38 ; total NO = 48

Author(s)	Year	Total Cites	#Cites of AFTER	Higher cites than after (Y/N)
ADAMS WJ	1970	7	38	**NO**
SUNDARARAJAN V	1970	5	26	**NO**
STONE JM	1971	0	74	**NO**
ARROW KJ	1971	45	26	YES
PAPANEK GF	1971	0	19	**NO**
ROTHSCHI.M	1971	72	17	YES
CONNOLLY M	1972	13	84	**NO**

QUIGLEY JM	1972	8	6	YES
SCHYDLOW.DM	1972	9	17	**NO**
PIERSON G	1972	0	0	**NO**
REPETTO R	1972	0	0	**NO**
SMITHIES A	1972	0	0	**NO**
GINTIS H	1972	27	56	**NO**
SELOWSKY M	1973	1	19	**NO**
SPENCE M	1973	695	21	YES
JORGENSO.DW	1973	7	53	**NO**
ROBERTS MJ	1973	2	6	**NO**
FELDSTEI.MS	1974	65	0	YES
MUSGRAVE RA	1974	20	26	**NO**
COOTER R, HELPMAN E	1974	21	6	YES
WEINSTEIN MC, ZECKHAUSER RJ	1975	68	110	**NO**
SPENCE M	1976	16	143	**NO**
LEVITT T	1976	2	4	**NO**
SPENCE M	1976	17	140	**NO**
CAVES RE, PORTER ME	1977	356	12	YES
BRINNER RE	1977	7	5	YES
AUERBACH AJ, PELLECHIO AJ	1978	7	55	**NO**
PRATT JW, WISE DA, ZECKHAUSER R	1979	84	5	YES
HARTMAN DG	1979	5	24	**NO**
AUERBACH AJ	1979	78	4	YES
SHAVELL S	1979	124	147	**NO**
AUERBACH AJ	1979	15	29	**NO**
FELDSTEIN M, HARTMAN D	1979	29	8	YES
WEINSTEIN MC, SHEPARD DS, PLISKIN JS	1980	58	6	YES
FRIEDMAN BM	1980	4	0	YES
FREEMAN RB	1980	163	20	YES
SACHS J	1980	59	201	**NO**
CLARK KB	1980	47	30	YES
LAZONICK W	1981	21	11	YES
LOONG LH, ZECKHAUSER R	1982	2	14	**NO**
BELL C, DEVARAJAN S	1983	4	14	**NO**
ABEL AB	1983	4	0	YES
SHAVELL S	1984	47	16	YES
GHEMAWAT P, SPENCE AM	1985	40	15	YES
MANKIW NG	1986	38	70	**NO**
RODRIK D	1987	21	228	**NO**
GREEN J	1987	30	94	**NO**
ESPINOSA MP, RHEE CY	1989	29	55	**NO**
STEIN JC	1989	145	270	**NO**
WEIL P	1990	109	16	YES
FRIEDMAN BM, WARSHAWSKY MJ	1990	57	48	YES
BARRO RJ	1991	1136	237	YES
DELONG JB, SUMMERS LH	1991	237	159	YES
WEITZMAN ML	1992	148	895	**NO**
WEITZMAN ML	1993	77	23	YES
SUMMERS L, GRUBER J, VERGARA R	1993	32	47	**NO**
LEAHY JV	1993	44	40	YES
MADRIAN BC	1994	60	24	YES
HINES JR, RICE EM	1994	63	85	**NO**
BEBCHUK LA	1994	33	191	**NO**
ELLISON G, FUDENBERG D	1995	76	0	YES
ADES AF, GLAESER EL	1995	60	10	YES
FELDSTEIN M	1995	5	178	**NO**
ISLAM N	1995	286	0	YES
LEVITT SD	1996	86	83	YES
KANE TJ, STAIGER D	1996	31	155	**NO**
BORJAS GJ, HILTON L	1996	53	81	**NO**
HOXBY CM	1996	45	76	**NO**
LAIBSON D	1997	287	61	YES
CUTLER DM, GLAESER EL	1997	123	17	YES
LOPEZDESILANES F	1997	28	109	**NO**
WEITZMAN ML	1998	29	44	**NO**
CUTLER DM, REBER SJ	1998	66	31	YES

GOLDIN C, KATZ LF	1998	107	41	YES
FOOTE CL	1998	21	27	**NO**
CAMPBELL JY, VICEIRA LM	1999	62	117	**NO**
BARRO RJ	1999	22	95	**NO**
LA PORTA R, LOPEZ-DE-SILANES F	1999	40	131	**NO**
WEITZMAN ML	2000	12	23	**NO**
ALESINA A, LA FERRARA E	2000	95	35	YES
HOXBY CM	2000	51	19	YES
LAIBSON D	2001	29	39	**NO**
HOXBY CM	2001	18	56	**NO**
ALESINA A, BARRO RJ	2002	28	72	**NO**
FRANKEL J, ROSE A	2002	72	4	YES
SAEZ E	2002	28	3	YES

The principal observations for the QJE European test (where Yes indicates discrimination)

[France+Spain+Italy+Germany+Switzerland+Belgium+Sweden+Holland]

TOTAL = 82 results; after exclusions = 27 results; total YES = 12 ; total NO =15

Author(s)	Year	Total Cites	#Cites of BEFORE	Higher cites than before (Y/N)
AUBAREDA J	1979	1	17	**NO**
SCHNEIDER F, POMMEREHNE WW	1981	38	56	**NO**
BEATO P	1982	8	11	**NO**
FITZROY FR, KRAFT K	1987	41	26	YES
PAGANO M	1989	61	11	YES
BARBOLLA R, CORCHON LC	1989	2	70	**NO**
DEWATRIPONT M	1989	55	29	YES
JAPPELLI T	1990	88	63	YES
FORGES F	1990	17	42	**NO**
DASPREMONT C, FERREIRA RD, GERARDVARET LA	1990	6	31	**NO**
DELBONO F, DENICOLO V	1991	7	15	**NO**
SAINTPAUL G	1992	44	19	YES
WALDMANN RJ	1992	60	70	**NO**
KIRMAN A	1993	91	53	YES
BERGLOF E, VONTHADDEN EL	1994	50	89	**NO**
TILMAN EHRBECK, ROBERT WALDMANN	1996	45	42	YES
NONNEMAN W, VANHOUDT P	1996	35	46	**NO**
ELLINGSEN T	1997	17	109	**NO**
GNEEZY U, POTTERS J	1997	35	19	YES
BOLTON P, ROLAND G	1997	83	109	**NO**
PAGANO M, ROELL A	1998	42	14	YES
GUISO L, PARIGI G	1999	39	14	YES
FEHR E, SCHMIDT KM	1999	509	35	YES
THESMAR D, THOENIG M	2000	11	72	**NO**
GUESNERIE R	2001	1	60	**NO**
MILESI-FERRETTI GM, PEROTTI R, ROSTAGNO M	2002	39	8	YES
ANDERSON S, BALAND JM	2002	10	13	**NO**

The principal observations for the Journal of Political Economy test on Chicago authors (where No indicates discrimination)

Total = 326 results; after exclusions = 104; total YES = 72; total NO = 32.

Author(s)	Year	Total Cites	#Cites of AFTER	Higher cites than after (Y/N)
LAFFER AB	1970	10	1	Y
PASHIGIAN BP	1970	9	0	Y
STIGLER GJ	1970	274	5	Y
FAMA EF	1971	49	25	Y
KESSEL R	1971	72	12	Y
FRIEDMAN M	1971	94	1	Y
SEITZ WD	1971	18	8	Y
NERLOVE M	1972	15	110	N
STIGLER GJ	1973	2	13	N
EHRLICH I	1973	492	74	Y
FAMA EF, MACBETH JD	1973	823	2888	N
FISCHER S, COOPER JP	1973	17	94	N
SCHULTZ TW	1974	7	140	N
BECKER GS	1974	140	5	Y
BENHAM L	1974	20	5	Y
BARRO RJ	1974	1,165	155	Y
LEWIS HG	1974	46	11	Y
STIGLER GJ, FRIEDLAND C	1975	53	108	N
POSNER RA	1975	370	0	Y

LUCAS RE	1975	353	**44**	**Y**
JOHNSON HG	1976	12	**2**	**Y**
HECKMAN JJ	1976	152	**51**	**Y**
BECKER GS, TOMES N	1976	153	**19**	**Y**
STIGLER GJ	1976	15	**56**	**N**
PASHIGIAN BP	1976	21	**72**	**N**
TOLLEY GS, WILMAN JD	1977	24	**62**	**N**
REID JD	1977	31	**18**	**Y**
LAZEAR E	1977	23	**19**	**Y**
EHRLICH I	1977	77	**22**	**Y**
TELSER LG, HIGINBOTHAM HN	1977	65	**23**	**Y**
HARBERGER AC	1978	59	**17**	**Y**
DRAZEN A	1978 1978	72	**13**	**Y**
LINNEMAN P	1978	12	**4**	**Y**
MUSSA M	1978	65	**19**	**Y**
MISHKIN FS	1979	12	**3**	**Y**
KIEFER NM	1979	12	**43**	**N**
CARLTON DW	1979	41	**12**	**Y**
LAZEAR EP	1979	430	**23**	**Y**
FAMA EF	1980	1,119	**35**	**Y**
LANDES EM	1980	16	**73**	**N**
REDER MW, NEUMANN GR	1980	88	**7**	**Y**

GREGORY N	1980	11	6	Y
LANDSBURG SE	1981	21	89	N
FRENKEL JA	1981	214	254	N
MISHKIN FS	1982	184	86	Y
MUSSA M	1982	116	21	Y
SINDELAR JL	1982	41	13	Y
MILLER MH, SCHOLES MS	1982	104	16	Y
KORMENDI RC, MEGUIRE PG	1984	42	29	Y
JARRELL G, PELTZMAN S	1985	91	11	Y
LAHAYE L	1985	11	54	N
LIEBOWITZ SJ	1985	59	21	Y
ZARNOWITZ V, LAMBROS LA	1987	78	4	Y
FRIEDMAN D	1987	10	0	Y
HARTZMARK ML	1987	20	33	N
FAMA EF, FRENCH KR	1988	416	7	Y
PASHIGIAN BP	1988	3	20	N
STOKEY NL	1988	102	73	Y
TOPEL R, ROSEN S	1988	73	3	Y
WERNERFELT B	1988	3	41	N
SHLEIFER A, VISHNY RW	1988	9	57	N
LAZEAR EP	1989	166	87	Y
DIAMOND DW	1989	145	35	Y
BULOW J,	1989	84	11	Y

ROBERTS J				
CAMERER C, LOEWENSTEIN G, WEBER M	1989	103	**37**	**Y**
TOWNSEND RM	1989	29	**30**	**N**
SUEN W	1989	19	**172**	**N**
CONSTANTINIDES GM	1990	326	**50**	**Y**
ROMER PM	1990	637	**253**	**Y**
SNYDER JM	1990	77	**38**	**Y**
TOPEL R	1991	213	**22**	**Y**
SAH RK	1991	29	**16**	**Y**
DIAMOND DW	1991	225	**194**	**Y**
SAH RK	1991	99	**10**	**Y**
IRWIN DA	1991	10	**3**	**Y**
HUBBARD RG, KASHYAP AK	1992	55	**29**	**Y**
KANDEL E, LAZEAR EP	1992	182	**145**	**Y**
FRIEDLAND C	1993	4	**4**	**N**
PELTZMAN S	1993	10	**3**	**Y**
WATSON MW	1993	87	**73**	**Y**
CHICAGO U, MURPHY KM, SCHEINKMAN JA	1994	**0**	**12**	**N**
GAREN JE	1994	73	**91**	**N**
IRWIN DA, KLENOW PJ	1994	91	**70**	**Y**
STIGLER GJ, STIGLER SM, FRIEDLAND C	1995	38	**170**	**N**

Telser LG	1996	3	**41**	**N**
Attanasio O, Davis SJ	1996	53	**74**	**N**
Debelle G, Lamont O	1997	19	**8**	**Y**
Diamond DW	1997	30	**5**	**Y**
Mulligan CB	1997	19	**50**	**N**
Young A	1998	106	**41**	**Y**
Philipson TJ, Becker GS	1998	21	**27**	**N**
Mulligan CB	1998	7	**50**	**N**
Levitt SD	1998	37	**35**	**Y**
Mulligan CB	1999	22	**22**	**N**
Chiappori PA, Salanie B	2000	44	**35**	**Y**
Cochrane JH, Saa-Requejo J	2000	35	**36**	**N**
Goolsbee A	2000	22	**37**	**N**
Peltzman S	2000	36	**18**	**Y**
Garicano L	2000	51	**12**	**Y**
Diamond DW, Rajan RG	2001	51	**3**	**Y**
Luttmer EFF	2001	39	**13**	**Y**
Duggan M	2001	34	**12**	**Y**
Vissing-Jorgensen A	2002	43	**12**	**Y**
Prendergast C	2002	64	**4**	**Y**