Statistical Science
1993, Vol. 8, No. 3, 310–330

Report of the Ad Hoc Committee on Double-Blind Refereeing

David Cox, Leon Gleser, Michael Perlman, Nancy Reid (Chair) and Kathryn Roeder

Abstract. This report of the Ad Hoc Committee on Double-Blind Refereeing was presented to the IMS Council in August 1992. It summarizes arguments for and against blinding referees to the identity of the authors and suggests an experiment to evaluate the process.

Key words and phrases: Refereeing, peer review, publication bias.

INTRODUCTION AND RECOMMENDATIONS

An ad hoc committee to consider double-blind refereeing was established in February 1991 by David Siegmund, then president of the Institute of Mathematical Statistics (IMS). The committee was formed at least partly in response to the report of the New Researchers Committee (Altman et al., 1991). The mandate to the committee was to review the literature and arguments in favor of and opposed to double-blind refereeing and summarize their relevance to IMS and its publications. It was stated that it would also be very interesting if an experiment might be suggested that the IMS could consider performing.

The committee reviewed much of the available literature and collected opinions on an informal basis from a haphazard sample of IMS members. Broadly speaking, the majority opinion among those senior enough to have had experience of the editorial process does not favour double-blind refereeing, and the majority opinion among those in possibly disadvantaged groups does favour double-blind refereeing. These views are summarized in more detail in the following sections.

This committee recommends that if a change to double-blind refereeing is contemplated that it not be implemented without first assessing the merits of double-blind refereeing by an experiment. Some details are provided in later sections indicating issues to be considered in implementing an experiment. We feel that the available results in the literature are sufficiently incomplete that a well-designed experiment has the potential to provide useful information on the refereeing process, even if the eventual decision of the IMS is not to change the refereeing method.

The committee also recommends that if an experiment is to be conducted, a principal investigator with experience in supervising clinical trials be appointed to oversee it. We recommend that a preliminary pilot study be performed prior to implementing the full experiment.

David Cox is Warden, Nuffield College, Oxford OX1 1NF, United Kingdom. Leon Gleser is Professor, Mathematics and Statistics Department, University of Pittsburgh, Thackeray Hall, Pittsburgh, Pennsylvania 15260. Michael Perlman is Professor, Statistics Department GN-22, University of Washington, Padelford Hall, Seattle, Washington 98195. Nancy Reid (chair) is Professor, Statistics Department, University of Toronto, 100 George Street, Toronto M5S 1A1, Canada. Kathryn Roeder is Associate Professor, Statistics Department, Yale University, Yale Station Box 2179, New Haven, Connecticut 06520.

SUMMARY OF VIEWS ON DOUBLE-BLIND REFEREEING

Double-blind refereeing means that the name of the author is not available to the referee and the name of the referee is not available to the author. While the suggestion has often been received that the editor and associate editors have the more important influence on fairness in publication, the person requesting referees' reports must know the authors' identities, if only to avoid asking authors to referee their own papers.

The following provides a list of the more cogent arguments received by the committee on the advantages and disadvantages of double-blind refereeing.

Advantages

- Double-blind refereeing allows the merits of a paper to be assessed without regard to characteristics of its author(s) such as rank, gender, institution, seniority, and so on. (Although the identity of the author may be guessed, it is argued that the effect of this is slight.)
- Double-blind refereeing ensures that the perception of decisions being based on such unrelated

characteristics is eliminated, which may indirectly benefit all its members.
- The amount of work involved is small, and there seems to be nothing to lose by implementing double-blind refereeing.
- The name of the author is not relevant for assessing the suitability for publication of a submitted manuscript.

Disadvantages

- Most editors and many referees feel that they provide extra advice to new researchers, particularly over presentation; this extra assistance will not be available under double-blind refereeing.
- It may be advantageous, particularly for new researchers, to publicize their prepublication work through the refereeing process.
- The name of the author is a relevant piece of information that makes the refereeing process more efficient, in that judgements about the likely impact of the work are legitimately influenced by the author's name, background and the importance of his or her previous work.
- Some distortion in the refereeing process may be introduced by referees trying to guess the author, sometimes of course incorrectly.
- There is some increased editorial burden in changing the cover sheet of the submitted paper to remove the authors' names. (This is a slight burden that would eventually be transferred to the authors, if double-blind refereeing became journal policy.)

Views on the merits of double-blind refereeing are quite varied and strongly held. Comments the committee received spanned a range of opinions from (loosely quoted) "double-blind refereeing is a waste of time" to "double-blind refereeing is the only fair refereeing process." The suggestion was also received by the committee that a more effective way to ensure the quality of referees' reports is to have referees sign their reviews. Preliminary analysis of the IMS survey questionnaire indicates strong support for double-blind refereeing in the IMS journals. (Of 1,141 responses analysed by July 6, 1992, 46% agreed or strongly agreed with the statement that the IMS should institute double-blind refereeing, 24% were neutral and 29% responded negatively.)

The New Researcher's Committee (NRC) strongly endorses double-blind refereeing in its recent report and again in its rejoinder to the discussion (Altman et al., 1991) and has been a major force behind the formation of the present committee. It seems likely that the NRC represents the majority opinion among new researchers, although support for double-blind refereeing is not unanimous among new researchers.

The publication committee in its report to the board in August 1991 was "negative but sympathetic," and this seems to be a majority view among those senior enough to have been involved in the editing process. An informal poll of some past editors of the *Annals* indicated little support for double-blind refereeing.

The present committee's view may be summarized as cautiously receptive to double-blind refereeing. We are not convinced that the benefits outweigh the disadvantages; but we are open to the possibility. We recommend that if a change in journal policy to double-blind refereeing is contemplated that an experiment be conducted to assess the merits of double-blind refereeing before any permanent change is made. Some general issues about the experiment are discussed in a later section, and some guidelines for the design are provided.

LITERATURE

A literature review is provided in the last section. As might be expected, although there are many published articles describing the possible existence of peer review bias, none can be regarded as definitive. There are also studies indicating moderate variability between referees evaluating the same paper.

The most widely quoted study is that by Peters and Ceci (1982). There is little agreement on the validity of the conclusions from this study, as evidenced by the disagreement expressed by a large number of discussants. The *Journal of the American Medical Association* has also conducted studies of double-blind refereeing. The most relevant of these was a randomized block design which compared the quality of reports received from both the editors' and authors' points of view. This study did not investigate association of differences between reports with gender, institution, seniority and so on.

OTHER JOURNALS' EXPERIENCE

The only statistics journals that we know of that have implemented double-blind refereeing are the *Canadian Journal of Statistics* and *Psychometrika*. The *Canadian Journal of Statistics* switched to double-blind refereeing in January 1991. Marc Moore was the editor during that period; the editor since January 1992 has been Lai Chan. The issue of double-blind refereeing was not studied systematically, either before or after the implementation. Marc Moore commented that the editorial burden in blinding the submissions was indeed very small, and his subjective impression was that the refereeing process was very little affected by the change in journal policy. There was an occasional objection to the policy, by a referee or an author, but these objections were not strong, and no submissions were withdrawn nor referee reports refused on these grounds. The current editor confirmed these views.

Technometrics allows authors the option of requesting double-blind refereeing; the editor Vijay Nair reported that a very small number of submissions take advantage of this option. Again, the subjective opinion of the editor was that this had essentially no effect on either the journal or the outcome for the papers in question.

EXPERIMENT: GENERAL

It is important to note that an experiment to evaluate double-blind refereeing does involve a substantial amount of work on the part of the editors and at least some members of the IMS. There is little available information to provide a guess at how much can be learned about the refereeing process even from a carefully designed experiment. Most editors feel that there is considerable variability among referees evaluating the same paper, particularly for papers that are not at the two quality extremes, although even this has not been systematically addressed. There is also the difficulty that reviews prepared as part of an experiment might not be representative of the reviews obtained under the usual conditions of operation of the journal – in particular, that reviewers who know their reports will be assessed might be more careful and thorough than usual.

On the other hand, we would probably agree as a discipline that a carefully designed experiment is the appropriate way to obtain information, and of course such an experiment has intellectual value in itself, aside from whatever impact it may have on the IMS journal policy.

Although the views of the members of the IMS, as received on an ad hoc basis by this committee, are widely varying on the merits of double-blind refereeing, there is little disagreement on the merits of an experiment to attempt to find out more about the refereeing process.

We recommend that, if an experiment is to be conducted, a principal investigator with experience in clinical trials be appointed to oversee the design and analysis. Some aspects of the design that will likely be relevant are outlined in the next section.

DESIGN

General

The following is intended to convey some of the issues that need to be considered in designing an experiment and to sketch some approaches to the problem.

It is essential to have the cooperation of the Editors in carrying out the experiment. Associate Editors should be encouraged but not compelled to take part. We suggest that a pilot study be carried out, involving one *Annals* and two Associate Editors, for a period of two months, at which time the data sheets and randomization instructions may need to be revised. We further recommend that after this the experiment be carried out for six months, on all three *Annals*.

Some papers that are submitted are not sent to Associate Editors by the Editor or not sent to referees by the Associate Editors. These papers will of course be handled in the usual way and will not form part of the experiment. In addition, some papers are specially solicited for the journal by either the Editor or the Associate Editors, and it is recommended that these papers also not form part of the experiment. Because *Statistical Science* has a large number of papers in the latter category, it might be excluded from the experiment.

The Associate Editor may need to serve as a second referee on a paper. This will presumably most often be because one of the two referees initially contacted does not produce a report within a reasonable length of time. In this case, the information about the paper should be collected as outlined below, with a note indicating that a complete randomized response was not obtained.

There may be papers involving more than two referees. The Associate Editor may serve as a third referee, either because she has a particular interest in the paper or because the two referees involved were in disagreement. In this case, as long as the randomization procedure was followed, the paper will still form part of the experiment. Some specific guidelines are needed in the event that the Associate Editor decides to send a paper to more than two referees. It may be necessary to find some way to keep records on the referees' reports for papers refereed in the usual way to assess the influence on quality of participation in the experiment. Once a decision on a paper has been reached, the editorial office involved should perhaps inform the referee of both the authors' names and the decision.

To give a flavour of the type and amount of effort that will be involved in an experiment, we provide the following suggestions for conducting the experiment, and a sketch of the data forms that might be used.

Editorial Office

- prepares each paper (that is to be sent to an Associate Editor) by putting information about authors, institutions, and acknowledgements on a separate cover sheet, so that the paper is "blind" when the cover sheet is removed
- records author information on the Data Sheet 1
- mails a sealed randomization envelope with the paper to the Associate Editor
- sends a letter to the author (along with the acknowledgement of receipt) indicating that the paper may be refereed blind

Associate Editor

- for each paper to be refereed by two referees, chooses the two referees and names them A and B
- opens the randomization envelope to find out which referee is to get the blind copy
- continues with the refereeing process in the usual way
- completes Data Sheet 2, which is returned to the Editor along with the usual correspondence

LITERATURE REVIEW

In the late 1970's and early 1980's, considerable criticism was directed at the peer review process. Both the National Science Foundation and the National Institutes of Health made studies of peer reviews of grant applications. Of greater interest for our purpose are the studies of the refereeing process that were made at that time. Articles reporting results from these studies are included in the bibliography of an article discussed by Peters and Ceci (1982), which will be mentioned later in this review. Although we have found some papers on the subject that were published after 1982, the great scope of the literature where relevant articles might appear has prevented a comprehensive search. We nonetheless believe that the present review gives an adequate summary of the issues and findings pertinent to the question of initiating and testing double-blind refereeing.

It will become apparent that the majority of our references come from the social and behavioral science literature. It is natural for these fields to be interested in the behavior of referees, but more important as a reason for the preponderance of articles from this literature is that journal acceptance rates in the social and behavioral sciences tend to be quite low (20–30%). The failure of authors to secure a share of scarce resources creates a discontent with the system. In contrast, even quite prestigious physics journals, such as *Physical Review*, accept approximately 80% of the articles submitted to them. A similar high acceptance rate occurs in some of the chemical journals. Consequently, there is less incentive to question the editorial practices of those journals.

The subsequent review covers the following issues: (a) the reliability of referees' ratings or judgements of articles, (b) the existence of status bias (favouring senior researchers or those at well-known institutions), (c) the existence of gender bias (favouring men over women) and (d) studies of the results of implementing double-blind refereeing. It is notable that we found no studies of racial or ethnic bias.

Reliability of the Refereeing Process

How closely do referees agree in their judgements about submitted articles? This is an important consideration when investigating bias in refereeing because bias and legitimate differences in judgement can be confounded.

In the behavioral sciences, between-rater agreement appears to be quite low. Perhaps the highest levels of agreement were reported by Scarr and Weber (1978). Of 89 articles submitted to the *American Psychologist*, there was perfect agreement between two referees on a 5-point rating scale for 57 of 89 papers; an additional 12 referees differed by only one rating category. For the same journal, Cicchetti (1980) found an intraclass correlation between referees' ratings of 0.54. A study by Cicchetti and Eron (1979) of 450 psychology manuscripts found an intraclass correlation of $r = 0.15$ (for a 7-point rating scale). Another 600 manuscripts for which the rating scale was not used yielded $r = 0.21$. In sociology, McCartney (1973) examined 300 manuscripts submitted to *Sociological Quarterly* and found that, on a 5-point scale, approximately one third of referee pairs were in perfect agreement and one third differed by one level. Other reports of reliability studies in the social and behavioral sciences are: Crandall (1978), Gottfredson (1978), Hendrick (1976, 1977), Mahoney (1977), McReynolds (1971), Scott (1974) and Watkins (1979).

Similar low levels of agreement have been found in the medical sciences. For example, Ingelfinger (1974) reported an intraclass correlation of less than 0.30 for reviewer ratings of 496 manuscripts for the prestigious *New England Journal of Medicine*. In the physical sciences, Lazarus (1982) remarks that no more than 15% of articles submitted to *Physical Review Letters* secure perfect agreement between referees, either on outright acceptance or outright rejection, and notes the importance of Editors or Associate Editors in resolving the resulting gray areas.

Given the space constraint pressures on journals in the behavioral, social and medical sciences, and the resulting pressure on referees to be critical and selective, one can conjecture that ratings of "reject" or "reject with encouragement" would be agreed upon most frequently by referees. This conjecture is confirmed by Cicchetti (1980).

In any case, at least for the journals covered by the above findings, it appears that few papers receive unanimous "accept" or "accept after minor revision" reviews. Those that do may have high quality or may simply be fortunate in having been reviewed by referees who were unusually uncritical. There may thus be some grounds for the belief held by many researchers that securing publication of their papers is largely the result of chance.

What about the validity of referees' judgements? A study by Garfunkel, Ulshen, Hamrick and Lawson (1990) tested the hypothesis that no important deficits would be identified on further review of accepted manu-

scripts. The authors selected the *Journal of Pediatrics* for the study. The new referees were unaware of the study or of the status of the papers. Although the referees did come up with additional criticisms, 80% of the papers were recommended for acceptance and the others were judged suitable, but not of high priority. (This study can also be viewed as measuring reliability of ratings of acceptance, because the actual quality of the manuscript is unknown.) On the other hand, Peters and Ceci (1982) submitted 12 previously published papers to the same psychology journals in which they had been published, but under new authors' names (from different institutions). The attempted plagiarism was spotted in only three cases; of the remaining nine papers, 8 were unanimously rejected by the referees!

Status Bias

Status bias refers to situations in which the best known researchers, senior researchers and/or researchers at prominent universities or research centers receive gentler than average treatment at the hands of referees, while unknown authors at less well known institutions receive harsher than average treatment. In the social sciences, such differential treatment of authors by their status is sometimes called the "Matthew Effect" (Merton, 1968), deriving from the biblical aphorism: "Unto every one that hath shall be given . . ., but from him that hath not shall be taken away even that which he hath." A related type of bias, commonly called "the old boys network," is the claim that prestigious individuals control access to funds and journal space for themselves and other prestigious researchers and may deny such resources to unestablished researchers. The possible existence of status bias or an "old boys network" underlay most of the calls for reform of the refereeing and grant review system in the late 1970's.

Status bias, indeed any bias, is difficult to demonstrate. Most studies that have attempted to verify status bias have inevitable flaws that are seized upon by those who prefer to keep their belief in an objective and fair peer system. As Peters and Ceci (1982) note, journal Editors are reluctant to cooperate in tests of bias of their referees even though their cooperation is essential for any well-designed experiment. In particular, the need to control the quality of manuscripts used for the test may require deceiving referees about the true status of the manuscripts they are reviewing. [See Fleiss (1982) for a contrary opinion.] This is viewed as unethical by many.

This was the case with the Peters and Ceci (1982) study mentioned previously. Because of lack of cooperation from Editors, they not only deceived the referees but also deceived the Editors. Apart from cosmetic changes in the title, introduction and the method used to display data, the papers submitted were identical to those previously published in the journals to which they were resubmitted. The major change was that the authors and affiliations were changed from those of well-known researchers in top psychology departments to unknown authors at institutions such as the "Tri-Valley Institute of Human Learning." The fact that of the 9 papers not spotted as plagiarisms, 8 were unanimously rejected is, given the variability one would expect in referee's opinions and the supposed publishable quality of the papers, a strong indication that the status of the supposed authors played a strong part in the referees' judgements. Some discussants of Peters and Ceci (1982) objected that the use of institutes of doubtful repute as locations for the supposed authors could legitimately create suspicion as to the validity of the data reported (falsification of data is not unheard of in science). Others noted that the design was incomplete, lacking (among other things) a control group to detect whether rejection was a consequence of poor quality control at the papers' original submission. Submission of low-quality papers under the names of well-known authors was also recommended. Finally, the small sample size of the study came in for criticism. Nevertheless, nearly all discussants expressed surprise and concern at the results.

Some of the physical scientists who discussed the Peters and Ceci (1982) study noted that the status of authors can provide important evidence of the quality of work not presented in detail in a paper (e.g., that the data were obtained correctly). See, for example, Yalow (1982).

Most other studies of status bias have been observational (uncontrolled). For example, Gordon (1980) studied 2,572 referees' reports on 619 articles submitted to well-regarded physical science journals in the United Kingdom. There was a significantly higher chance of a favourable evaluation ($p < 0.000006$) when both author and referee shared membership in groups of similarly high status. Gordon attributed this difference to a higher level of consensus on research beliefs and paradigms within groups of similar status than across such groups, but also admitted the possibility of prejudice due to status. He also noted that referees from lower status institutions did not exhibit significant institutional bias. Mahoney, Kazdin and Konigsberg (1978) found that manuscripts citing as "in press" articles under the author's name were more favourably reviewed than when these same "in press" articles were cited under the name of another researcher (indicating the advantage of reminding referees that the author has previously survived the peer review process). Rowney and Zenisek (1980) queried reviewers about the criteria they used in evaluating manuscripts and found that "justifiably strong reputation" of the author was frequently mentioned. Other related studies include

Beyer (1978), Bowen, Perloff and Jacoby (1972), Cicchetti and Eron (1979), Cole and Cole (1972), Crane (1967), Goodstein and Brazis (1970), Mahoney (1977), Merton (1968), Oromaner (1977), Pfeffer, Leong and Strehl (1977), Tobach (1980), Yoels (1974), Yotopoulos (1961), Zuckerman (1970) and Zuckerman and Merton (1973).

Note: The power of name recognition has also been demonstrated by Ross (1982). He submitted an award-winning novel (*Steps* by Jerry Kosinski, 1968) to fourteen major publishing houses under a pseudonym. No one recognized the work, and it was rejected by all of the publishers, including the original publisher.

Gender Bias

Only a small number of studies have been directed toward demonstrating the existence of gender bias. Anecdotally, Horrobin (1982) mentions that after women members of the Modern Language Association complained that gender bias was preventing them from publishing their work, the Association instituted a double-blind refereeing system; the result was a markedly higher rate of acceptance of articles written by women. Moore (1978) studied published reviews of books and found that reviewers gave more favourable reviews to their own sex than to members of the opposite sex; interestingly, both sexes were more critical of books authored by women.

Goldberg (1968) tested the hypothesis that both men and women value the professional work of men more highly by having women subjects evaluate supposedly published journal articles on linguistics, law, art history, education, dietetics and city planning. Half the women saw a male author's name (John T. McKay) and half saw a female author's name (Joan T. McKay). Results indicated that women rated the articles (even those considered sex-appropriate for women) more favourably when John was the alleged author. Paludi and Bauer (1983) extended this experiment to include 180 male and 180 female students as reviewers. They also added an author named J. T. McKay. Their results further supported the hypothesis of gender bias: J. T., whom many readers thought was female, rated somewhat lower than John, but higher than Joan. [See also Levenson et al., (1975).]

Double-Blind Refereeing

Double-blind refereeing appears to have been widely adopted by the behavioral and social science journals. Although Editors of double-blinded journals state opinions about the success of blinding based on their personal experience, this is generally not supported by controlled experimentation. The former Editor of the *Journal of Comparative and Physiological Psychology*, Garth J. Thomas, states: "Such evidence as I know indicates that such procedures do not make much difference in terms of actual administrative decisions (publish or reject). However, blind reviewing may be a desirable public relations maneuver as a reaction to the apparently increasing belief among scientists that the editorial process is shot through with bias and injustice" (Thomas, 1982). Sandra Scarr, who edited *Developmental Psychology* and served as Associate Editor of *American Psychologist*, claims (Scarr, 1982) that for both of these blind-refereed journals more than two-thirds of the referees appear to be genuinely blind to the authors' identity "as evidenced by the frequent tell-tale mistakes they make about the experience, gender and presumed theoretical positions of the authors. Young reviewers mistake senior investigators for novices, to whom they lecture; many reviewers refer to authors as 'he' or 'he/she' or 'they,' inappropriately; and some reviewers argue for a more extreme version of the author's own theoretical position, as though the author might need coaxing." She notes that blind refereeing sometimes has a detrimental effect by casting doubts on the competence of authors with well-established credibility, forcing them to spell out their methods in excruciating detail, but she claims it is "manifestly more fair to lesser known investigators in lesser known institutions." Rosenblatt and Kirk (1980) found that only 15–30% of referees for the *Journal of Social Service Research* could correctly identify the author(s) of manuscripts they had reviewed. In contrast, Robert K. Adair, the former Editor of *Physical Review Letters*, asserts (Adair, 1982) that blind refereeing would not work in physics journals, based on unpublished studies that "suggest to us that almost 80% of the authors of letters submitted to *Physics Review Letters* can be identified by referees competent in the narrow field of the submitted paper."

In research reported by McNutt, Evans and Fletcher (1990), each of 127 consecutive manuscripts submitted to the *Journal of General Internal Medicine* was sent to two reviewers; one was randomly chosen to receive the manuscript with the author and institution removed. After the reviews were returned, an Editor who was blinded to both the author and the blinding status of the referees evaluated the quality of the reviews. The quality was evaluated as the sum of a number of factors which took into account the Editor's point of view and the author's point of view. Over all, the blinded reports were graded to be of a higher quality (3.5 vs. 3.1 on a 5-point scale; p-value = .007). Components of the evaluation from the Editor's point of view included the following: (1) Did the reviewer give appropriate attention to the importance of the question studied? (2) Did the reviewer target key issues? (3) Did the reviewer clearly identify strengths and weaknesses in the study's methods? (4) Did the reviewer make constructive comments about the quality of writing and

presentation of the data? The blinded reviewer scored significantly higher in areas (1) to (3). Also, components of the evaluation from the author's point of view included (1) thoroughness, (2) constructiveness, (3) fairness, (4) courteousness and (5) knowledge in the manuscript's content area. The referees were graded on these points by both a blinded Editor and a blinded author. From the Editor's point of view, the blinded reviewer scored significantly higher on point (5) only. From the author's point of view there were no significant differences in the quality of the reviews. [Note: Of relevance to this point is the study done by Garfunkel, Lawson, Hamrick and Ulshen (1990), who asked 90 authors who submitted papers to *Journal of Pediatrics* to evaluate the quality of the reviews. The 60 rejectees did not rate the evaluations significantly lower than did the 30 acceptees.]

The referees also evaluated the quality of the manuscripts on a number of points. Blinded referees usually gave lower scores on most points, but the only significant difference occurred for methods. The referees, however, on the average did not differ in their recommendations regarding whether or not to accept the manuscripts.

No effort was made in this study to assess the influence of the author's institution, seniority, gender, etc. on the refereeing process. One difficulty with the study is that the referees receiving the blinded copy of the manuscript would have been aware that they were part of an experiment and may in consequence have been more careful with their reports.

DISCUSSION

As can be seen from the above literature review, there have been few well-designed studies of the refereeing process and even fewer studies of double-blind refereeing. In digesting the studies that have been done, it must be kept in mind that not all of the journals studied are similar to those of the IMS. That this is an important consideration can be seen by contrasting assertions made above about the behavioral and social science journals with those made about the journals of the physical sciences. It would seem that our journals share with the social and behavioral sciences the high journal rejection rates, with the unhappiness among authors that this produces. On the other hand, we like to think that our field more closely resembles the physical sciences in terms of having widely accepted paradigms of good research. Of the journals (and research areas) for which studies have been made of the refereeing process, it may be that the closest comparisons can be made to those in the medical area. Nevertheless, as statisticians we should be interested in providing good examples to other fields of how to run such studies and should agree that the best evidence about refereeing practices in statistics must come from studies of our own journals.

REFERENCES

ADAIR (1982). Discussion of "Peer-review practices of psychological journals," by D. P. Peters and S. J. Ceci. *The Behavioral and Brain Sciences* **5**.

ALTMAN, N. (CHAIR), BANKS, D., CHEN, P.-Y., DUFFY, D., HARWICK, J., LÉGER, C., OWEN, A. and STUKEL, T. (1991). Meeting the needs of new statistical researchers. A report by the New Researcher's Committee of the IMS. *Statist. Sci.* **6** 163–174.

ALTMAN, N., ANGERS, J. F., BANKS, D., DUFFY, D., HARDWICK, J., LÉGER, C., MARTIN, M., NOLAN, D. (CHAIR), OWEN, A., POLITIS, D., ROEDER, K., SRIRAM, T. N., STUKEL, T. and YING, Z. (1992). Rejoinder to Readers' comments to the New Researchers' Report. *Statist. Sci.* **7** 265–266.

BEYER, J. (1978). Editorial policies and practices among leading journals in four scientific fields. *Sociological Quarterly* **19** 68–88.

BOWEN, D. D., PERLOFF, R. and JACOBY, J. (1972). Improving manuscript evaluation procedures. *American Psychologist* **27** 221–225.

CICCHETTI, D. V. (1980). Reliability of reviews for the *American Psychologist*: A biostatistical assessment of the data. *American Psychologist* **35** 300–303.

CICCHETTI, D. V. and ERON, L. D. (1979). The reliability of manuscript reviewing for the *Journal of Abnormal Psychology*. In *Proceedings of the Social Statistics Section* 596–600. Amer. Statist. Assoc., Washington, D.C.

COLE, J. R. and COLE, S. (1972). The Ortega hypothesis: Citation analysis suggests that only a few scientists contribute to scientific progress. *Science* **1978** 368–374.

CRANDALL, R. (1978). Interrater agreement on manuscripts is not so bad! *American Psychologist* **33** 623–624.

CRANE, D. (1967). The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *American Sociologist* **32** 195–201.

FLEISS, J. L. (1982). Deception in the study of the peer-review process. *The Behavioral and Brain Sciences* **5** 210–211.

GARFUNKEL, J. M., LAWSON, E. E., HAMRICK, H. J. and ULSHEN, M. H. (1990). Effect of acceptance or rejection on the author's evaluation of peer review of medical manuscripts. *Journal of the American Medical Association* **263** 1376–1378.

GARFUNKEL, J. M., ULSHEN, M. H., HAMRICK, H. J. and LAWSON, E. E. (1990). Problems identified by secondary review of accepted manuscripts. *Journal of the American Medical Association* **263** 1369–1370.

GOLDBERG, P. A. (1968). Are women prejudiced against women? *Transactions* **5** 28–30.

GOODSTEIN, L. D. and BRAZIS, K. L. (1970). Credibility of psychologists: An empirical study. *Psychological Reports* **27** 835–838.

GORDON, M. D. (1980). The role of referees in scientific communication. In *The Psychology of Written Communication* (J. Hartley, ed.) 263–275. Kogan Page, London.

GOTTFREDSON, S. D. (1978). Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgements. *American Psychologist* **33** 920–934.

HENDRICK, C. (1976). Editorial comment. *Personality and Social Psychology Bulletin* **2** 207–208.

HENDRICK, C. (1977). Editorial comment. *Personality and Social Psychology Bulletin* **3** 1–2.

HORROBIN, D. F. (1982). Peer review: A philosophically faulty concept which is proving disastrous for science. *The Behavioral and Brain Sciences* **5** 217–218.

INGELFINGER, F. J. (1974). Peer review in biomedical publication. *American Journal of Medicine* **56** 686-692.

LAZARUS, D. (1982). Inter-referee agreement and acceptance rates in physics. *The Behavioral and Brain Sciences* **5** 219.

LEVENSON, H., BUFORD, B. BONNO, B. and DAVIS, L. (1975). Are women still prejudiced against women? A replica and extension of Goldberg's study. *Journal of Psychology* **89** 67-71.

MAHONEY, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research* **1** 161-175.

MAHONEY, M. J., KAZDIN, A. E. and KONIGSBERG, M. (1978). Getting published: effects of self-citation and institutional affiliation. *Cognitive Therapy and Research* **2** 69-70.

MCCARTNEY, J. L. (1973). Manuscript reviewing. *Sociological Quarterly* **14** 290, 444-446.

MCNUTT, R. A., EVANS, A. T. and FLETCHER, S. W. (1990). The effects of blinding on the quality of peer review. *Journal of the American Medical Association* **263** 1371-1375.

MCREYNOLDS, P. (1971). Reliability of ratings of research papers. *American Psychologist* **26** 400-401.

MERTON, R. K. (1968). The Matthew Effect in science. *Science* **159** 56-63.

MOORE, M. (1978). Discrimination or favoritism? Sex bias in book reviews. *American Psychologist* **33** 936-938.

OROMANER, M. (1977). Professional age and the reception of sociological publications: A test of the Zuckerman-Merton hypothesis. *Social Studies of Science* **7** 381-388.

PALUDI, M.A. and BAUER, W. D. (1983). Goldberg revisited: what's in an author's name. *Sex Roles* **9** 387-390.

PETERS, D. P. and CECI, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again (with discussion). *The Behavioral and Brain Sciences* **5** 187-255.

PFEFFER, J., LEONG, A. and STREHL, K. (1977). Paradigm development and particularism: Journal publication in three scientific disciplines. *Social Forces* **55** 938-951.

ROSENBLATT, A. and KIRK, S. A. (1980). Recognition of authors in blind review of manuscripts. *Journal of Social Service Research* **3** 383-394.

ROSS, C. (1982). Rejecting published work: similar fate for fiction. *The Behavioral and Brain Sciences* **5** 236.

ROWNEY, J. A. and ZENISEK, T. J. (1980). Manuscript characteristics influencing reviewers' decisions. *Canadian Psychology* **21** 17-21.

SCARR, S. (1982). Anosmic peer review: a rose by another name is evidently not a rose. *The Behavioral and Brain Sciences* **5** 237-238.

SCARR, S. and WEBER, B. L. R. (1978). The reliability of reviews for the *American Psychologist*. *American Psychologist* **33** 935.

SCOTT, W. A. (1974). Interreferee agreement on some characteristics of manuscripts submitted to the *Journal of Personality and Social Psychology*. *American Psychologist* **29** 698-702.

THOMAS, G. J. (1982). Perhaps it was right to reject the resubmitted manuscripts. *The Behavioral and Brain Sciences* **5** 240.

TOBACH, E. (1980). ". . . that ye be judged." In *An Evaluation of the Peer Review System in Psychological Research* (J. Demarest, Chairman). Open Forum at the American Psychological Convention, Montreal.

WATKINS, M. W. (1979). Chance and interrater agreement on manuscripts. *American Psychologist* **34** 796-797.

YALOW, R. S. (1982). Competency testing for reviewers and editors. *The Behavioral and Brain Sciences* **5** 240-241.

YOELS, W. (1974). The structure of scientific fields and the allocations of editorship on scientific journals: Some observations on the politics of knowledge. *Sociological Quarterly* **15** 264-276.

YOTOPOULOS, P. A. (1961). Institutional affiliation of the contributors to three professional journals. *American Economic Review* **51** 665-670.

ZUCKERMAN, H. (1970). Stratification in American science. *Sociological Inquiry* **40** 235-257.

ZUCKERMAN, H. and MERTON, R. (1973). Patterns of evaluation in science: Institutionalization, structure and functions of the referee system. In *The Sociology of Science* (N. Storer, ed.). Univ. Chicago Press.