the two arms. The former goal will be accomplished by assessing referee compliance with the blinding procedures. Too high a rate of referee refusal, particularly in the blind-review arm of the study, would argue against implementation of the study on a larger scale. Also arguing against implementation would be a high percentage of correct guesses of authorship by the blinded referees. Estimates of rater variability in large part will determine the sample size required for the full study. Additional goals of the pilot study will be to estimate the distribution of submitted manuscripts by prestige of the authors, prestige of the institutions and by gender and country of origin of the authors to determine if sufficient numbers of manuscripts will be available in selected categories to do subset analyses in a full study.

For this pilot study, it is recommended that only one of the IMS journals participate. *The Annals of Statistics* receives approximately 400 manuscripts a year, 90% of which are forwarded to the AE's for review. The remaining papers have either been solicited by the Editor or are manuscripts whose content and/or length are deemed inappropriate for the Journal. Thus, each month approximately 30 manuscripts are received by the 24 or so AE's. During this pilot, the letter acknowledging receipt of manuscripts would include a statement that the pilot study was being conducted. Consent to participate in the pilot would be implied by failure to withdraw the manuscript.

As an initial estimate of agreement between reviewers, we propose measuring percent agreement, in which referee ranking is categorized as either accept (or tentatively accept) or reject (or tentatively reject). Within each arm of the study, we would estimate the rate of agreement. Based on 100 pairs of reviewers for each arm, the precision of the estimated rate of agreement would be at worst $\pm 10\%$. This is a conservative estimate, based on assuming the true rate to be .5. One hundred or more manuscripts would also allow estimation of the distributions of author and institution characteristics with similar precision. The actual number of pairs available will be dependent on the refusal rate of proposed referees, which is itself a rate for which an estimate is sought. We propose that all eligible manuscripts submitted to the journal within a 4–6-month period be "subjects" for this pilot study. With an additional 4–6-month waiting period for submission of referee reports, it is anticipated that at least 100 complete review pairs would be obtained by the end of one year.

While we feel that this pilot study provides a practical model for evaluating the feasibility of studying blinded refereeing, there remain some problems that this design will not solve. This study focuses on evaluating biases at the referee level, but it does not provide a mechanism for studying potential biases by the AE's, who are ultimately responsible for weighing the validity of the referee reports.

7. EVALUATION OF THE PILOT STUDY

If the rate of referee refusal, or the rate of correct identification of authorship by blinded referees is not too high, then a full study may be deemed feasible, and estimates of variability will be obtained for sample size projections, based at least in part on variance components from an analysis of variance model for the 1–4 scoring scheme. The decision to proceed with the full study will be made by the IMS Council and the editorial boards of the journals, using the estimated rates, the projected sample sizes necessary to address the usefulness of blinded refereeing in important subsets, and other factors. A report on the implementation and results of the pilot study might be presented in *Statistical Science*. If the decision were made to proceed with the full study, an announcement could be made in the journals to outline the protocol to be followed for the experiment.

REFERENCES

POCOCK, S. J. (1983). *Clinical Trials: A Practical Approach*. Wiley, Chichester.

POCOCK, S. J. and SIMON, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31** 103–115.

Comment

L. Billard

May I first thank the committee members who assembled these reports for this contribution to the integrity of the scientific publication process of our discipline in general and the IMS journals in particular. The issue of double-blind refereeing today is one fraught with emotional overtones both rational and irrational, often subconsciously culturally based, and so is difficult for many of us to resolve equitably no matter how well intentioned. Thus, the Reid Committee can be congrat-

L. Billard is University Professor, Department of Statistics, University of Georgia, Athens, Georgia 30602-1952.

ulated for its efforts to gather both facts and views (not necessarily the same thing) in as balanced a manner as possible and the Crowley Committee for its elegant proposal for an experiment. My remarks here will be limited to two headings. The first is to draw attention to some additional studies available in the literature beyond those covered in the reports. These may further muddy the waters or provide some clarification depending on one's philosophical viewpoint! Second, comments will be made on a few specific points raised in the reports themselves.

The Reid Committee report provides a reasonably comprehensive review of the research literature of studies prior to 1982 and in the social and behavioural sciences, with a primary emphasis, not surprisingly perhaps, on those studies relating to the refereeing process itself and to status bias. There is, however, a plethora of studies in recent years (since 1982) as well as a growing body of literature concerned with disciplines outside of the social sciences in general (though not unfortunately with the mathematical sciences directly). We single out briefly a few of these. More complete details can be found in Billard (1991, 1992, 1993, 1994). Most of these studies deal with the perceived effect of gender. However, based on the (small number of) common points of reference, it is likely that conclusions relating to gender would also be applicable when dealing with institutional status and with academic or research age of the investigator.

That women average fewer publications than men has been suggested by several studies ranging from Astin (1973) over all disciplines, to Fish and Gibbons (1989) for economics. Attempts to explain this phenomenon have typically disclosed subtle discriminations of many dimensions and not just that which might be present in the refereeing process. For example, Michelson (1989) and Cole and Zuckerman (1984) suggested this may be a consequence of a lack of rewards to women as compared to those enjoyed by men for comparable work. Astin and Bayer (1973) showed that at the time of the Astin (1973) study, women were mostly appointed to teaching, as distinct from research, positions thus providing an explanation for an apparent reduced rate of publication. This is consistent with Blackburn, Behyman and Hall (1978), who observed that the differing rates were due to situational factors, at least as far as the biological, physical and social sciences and the humanities were concerned. Along the same lines, Freeman (1977) concluded the difference was partly due to discipline variation. Over, Over, Meuwissen and Lancaster (1990) dealing with graduates in psychology, concluded that after controlling for various factors such as impact of doctoral advisor, women and men graduates published at comparable rates. Likewise, Persell (1983) controlled for institutional affiliation and concluded that at least in education, the rates of publication were the same for both women and men. Thus, the preponderance of evidence based on these studies might suggest there is no essential bias in the refereeing process.

However, in contrast to these studies are those which showed that work done by women is perceived to be of a lower quality than work done by a man. These run the gamut of the relatively recent Davis and Astin (1987) work to the oft-quoted earlier one by Fidell (1970), in which identical vitae were sent to heads of departments (of psychology), with offers on average at the Associate Professor level being proffered to those purporting to be men but only at the Assistant Professor level for those believed to be women. [This phenomenon still appears to prevail today but with salary levels; see Billard (1992).] Perhaps the most compelling investigation into this phenomenon is the Paludi and Bauer (1983) study mentioned in the Reid Report itself. In that study, reviewers were asked to rate the papers on a scale of 1 (being tops) to 5. Interestingly, both male and female reviewers gave papers believed to be written by a woman (Joan T. McKay) an average rating of 3.0. However, men gave those papers believed to be written by a man (John T. McKay) an average 1.9 rating, whereas women gave the John McKay papers an average rate of 2.3. Given the low overall acceptance rate for submissions to statistical journals, the impact of the difference between a 1.9/2.3 and a 3.0 rating would be substantial. While it was not an intent of this study to conclude so, the fact that women reviewers also gave male-authored papers a better rating and the same 3.0 rating to the Joan T. McKay papers as did the men reviewers supports the contention that biases truly are cultural and subconscious, that they do exist no matter how well intentioned we all most assuredly are and no matter how much we want to believe they do not exist.

An equally convincing study is that of the Lefkowitz (1979) report of the Modern Language Association (MLA) experience. Unlike the statistical societies (whose policies on this issue are not in question here), contributed papers at MLA meetings had first to survive a review stage before acceptance to be read. Prior to 1974, these papers were refereed with the author's name intact. In 1974, double-blind refereeing was tried with the effect that the number of women and of new investigators having papers accepted doubled from previous years. This number doubled again when repeated in 1975, until, by 1978, the proportion of acceptances among women and new researchers was comparable to that for men. The MLA Board subsequently decided in 1979 to use double-blind refereeing for all their publications. This may explain why the humanities of all disciplines experiences the smallest gap between the external indicators of career progress for men and women, although this could also be due to the larger numbers of women serving in the field (Billard, 1992). This MLA study is one of the few

that includes a focus on new investigators. While one must be careful of course not to infer too much from this single study, the fact that new investigators experienced the same fate as did women investigators suggests that some of the other results dealing only with gender issues likely pertain here too. Likewise, the Peters and Ceci (1982) study, well described in the Reid Report itself, albeit small, is sufficiently suggestive that many of the gender study results also may apply when translated into institutional status terms.

Therefore, returning to the Reid Committee Report itself, we see that the advantages listed under the section heading "Summary of views on double-blind refereeing" are in one form or another essentially endorsed by some of the literature study conclusions. While there has been no definitive study on this effect for statistical publications, even if the proposed experiment, or a different study, showed biases did not exist, the second listed advantage relating to "the *perception* of . . ." is important, and this should have a real and positive impact were a decision to double-blind referee publications to be made. It was harder, at least for me, to understand the logic of the listed disadvantages, especially as some appeared to be synonymous with a counterpart under the list of advantages. The last two are in my opinion trivial and of little consequence. It was not clear to me why the actions designed to assist new researchers, as in the first two listed disadvantages, are precluded. For example, the extra advice given the new researchers can still be offered but after the reports of the referees are completed at which stage the editor can pass on the added information that this author is a new researcher.

My major concern, however, pertains to the third so-called disadvantage in which it is suggested that the name of the author is relevant especially in assessing the impact of the work to be legitimately influenced by the author's record. This reason as stated if I have understood it correctly, provides a reason *for* double-blind refereeing, not one *against* it. It seems to me that it should be the work itself, and not the reputation of the author, which influences. . . . As statisticians, one of our maxims is that the data should speak for themselves, so likewise should we let the work speak for itself without undue influence from outside pressures such as those suggested here. As one ponders on this "disadvantage," it is instructive to recall Persell's (1983) study in which he established a so-called quality index. After controlling for number of publications, citations, rated quality of research and so on, he showed that for women, this index was negatively related to the quality of her work but it was positively related for men. Later, Chamberlain (1988), and earlier, Cole (1979) drew similar conclusions, thus supporting the Persell index theory results. Davis and Astin (1987) also showed that women's work is perceived to be of lower quality. Thus, if it is deemed to be important that an author's impact and reputation are relevant components of the review process, as this third listed "disadvantage" suggests, the perceived lower status of gender authorship (and similarly, the institutional address, and the as-yet-unestablished reputation of the younger researcher) is but perpetuated. It is therefore only to be expected that senior established researchers will tend to seek the status quo, being less inclined to want to move to double-blind refereeing, while new (and also women and researchers in lower status named institutions) researchers will tend to prefer that double-blind refereeing be introduced.

Nevertheless, my inherent trust in my own colleagues, as people not just as statisticians, would be vindicated if the double-blind experiment did in fact demonstrate that refereeing bias did not exist in our own journals. While I agree with the apparently relative consensus opinion that it should be done, and while I agree with the Crowley Committee that the experiment can be easily sabotaged by the refusal of reviewers to participate as double-blind reviewers, I would prefer to take the view that the intrinsic integrity of our profession's colleagues will ensure that no such undermining of the experiment will occur. In the general clinical trial setting, the statistician relies heavily on the cooperation of the trial's participants; presumably when the statistician becomes the participant, should not the experimenter (i.e., the statistical profession) expect nothing less than full cooperation? As the reports implied, our profession, of all disciplines, has the obligation to show how such an experiment should be done. If the results show there are biases in the refereeing process, then we surely want to know this so that we can seek ways to correct the inherent inequities. If there are no biases present, it is also comforting to learn that we have risen above the usual cultural influences of our society and are true to our profession as unbiased (albeit still variable no doubt) statisticians. The Crowley Committee's version of the experiment should be endorsed and implemented!

ACKNOWLEDGMENTS

Partial support from the National Science Foundation and the Office of Naval Research is gratefully acknowledged. The work was performed at the Isaac Newton Institute of Mathematical Sciences, Cambridge, United Kingdom.