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.