*Statistical Science*
1993, Vol. 8, No. 3, 318–330

# Report of the Ad Hoc Committee on Design of an Experiment on Double-Blind Refereeing

**Jacqueline Benedetti, Stephanie Green, Mei-Ling Lee and John Crowley (Chair)**

*Abstract.* The design of a pilot study to assess the feasibility of conducting a full-scale trial of blind refereeing for the Institute of Mathematical Statistics' journals is proposed. Issues related to the choice of outcome measure, methods of randomization, number of reviewers and duration of the pilot study are discussed. Consideration is also given to the identification and collection of author characteristics hypothesized to affect publication decisions.

*Key words and phrases:* Blind refereeing, study design, pilot study, clinical trials.

## 1. INTRODUCTION

In February 1991, the Institute of Mathematical Statistics (IMS) established an ad hoc committee to review the literature regarding double-blind refereeing and to consider whether this issue had relevance to the IMS and its publications. The Reid Committee recommended that if a change to double-blind refereeing was to take place that an experiment first be conducted, overseen by a group with experience in conducting and supervising clinical trials. Further, the committee recommended that a pilot study be carried out prior to the implementation of a full experiment.

As part of their report, the committee suggested a general design for the experiment that involved random assignment of two reviewers for each paper, one reviewer who would be aware of the identity of the authors and one who would not. Several draft forms were included, which chiefly involved assessment by the Associate Editor (AE) of the quality of the two referee reports. The implication was that the chief outcome measure of the experiment involved this quality assessment.

In response to this report, one of us (J.C.) was appointed to form a design committee whose main charge would be to assess the feasibility of conducting a full-scale study of the use of blind refereeing for the journals. As part of this mission, this committee developed a study design based on the use of referee rating as the primary outcome measure as an alternative to the design proposed in the Reid Report. Details of this design are discussed below.

In considering an appropriate design for this experiment, several key issues were addressed. These included the number of reviewers assigned to each manuscript, the method of randomization, the choice of an outcome measure and the identification of author characteristics that are of interest to assess with respect to outcome. We discuss each of these separately.

## 2. NUMBER OF REFEREES

A standard procedure for the IMS journals is to assign two reviewers for each manuscript. If the paper is of particular interest to the AE, if the two reviewers disagree about the manuscript or if one of the initial two fails to produce a report in a reasonable time period, the AE may also review the paper. For this experiment, it is recommended that four reviewers be selected, two blinded and two unblinded with regard to author and institution identity. This design would help reduce variability due to reviewer and would also help reduce the problem associated with missing data due to refusal or delay by one of the referees. While it is recognized that this design may burden the pool of potential referees, and possibly increase the workload of the authors in requiring responses to four sets of comments, it is felt that it is necessary, at least for the pilot study, to assess variability in referee rating.

In practice, potential referees who decline participation are usually replaced. For the purposes of this study, it was felt that attempting to replace referees who declined would complicate the experiment, since the study assignment of the refusing author would need to be known to the AE, and this might introduce

*Jacqueline Benedetti is Associate Professor, Department of Biostatistics, University of Washington, Seattle, Washington 98195. Stephanie Green and John Crowley are members of the Program in Biostatistics, Fred Hutchinson Cancer Research Center, Seattle, Washington 98104. Mei-Ling Lee is Assistant Professor, Channing Labs, Harvard University, Cambridge, Massachusetts 02142.*

bias in the choice of the new referee. There are two ways around this problem. One would be to have the randomization assignment known to the editorial staff, but not the AE. The other would be for the AE to have a list of alternate reviewers prepared ahead of time, so that the choice of whom to substitute would have been made prior to the randomization assignment. Neither of these alternatives is without problems. In the former case, the plan creates extra work for the journal staff, and it is probably not practical. Moreover, identification of additional referees for each article is likely to be too difficult to be realistic. The refusal rate itself would be viewed as one measure of the success of a blinded review process. That is, it would be of interest to assess whether refusal was higher among those assigned to the blind review compared to those who were aware of author identity.

## 3. RANDOMIZATION

In a multicenter clinical trials setting, the most common randomization scheme involves a centralized randomization office, which is contacted for patient assignment (Pocock, 1983). Patients are typically randomized according to some dynamic allocation scheme which balances on stratification factors of interest (Pocock and Simon, 1975). In this current setting, we have a design for which we wish to balance only by manuscript – that is, among four referees, two are to be assigned to blinded review, two to author knowledge. A common alternative to the central system is to employ sealed envelopes, which contain study assignment and which are to be opened once a set of referees has been identified for a particular manuscript. Thus, the four referees are assigned alphabetically to unique identifiers such as A, B, C and D. Once assigned, the envelope is opened to determine which two of these identifiers are assigned to the blinded group and which two to the unblinded. This randomization scheme moves the treatment assignment to the local level. However, the centralized scheme removes any potential temptation by an AE to alter the assigned "treatment" to a particular referee. Thus, it is recommended that centralized randomization be utilized in the pilot study, with reassessment prior to implementation of the full experiment.

## 4. OUTCOME MEASUREMENT

The current practice in the journals is for the referees to summarize their recommendations on a four-point scale: 1 = accept; 2 = tentatively accept; 3 = tentatively refuse; 4 = refuse. This rating scheme is proposed as the primary outcome measure for each referee. This measure is preferable to the AE assessment of referee report quality, which is itself subject to individual biases for which there are no adequate controls. Although other outcomes could be considered (e.g., ultimate acceptance of the paper, or referee responses to questions concerning style, innovation, applicability, etc.), the accept/reject scores provided by the referees seem most appropriate. The most direct effect of knowledge of author and institution is on the referees, and the most important information used by the AE's in determining ultimate acceptance is the referee rating. Among the comparisons of interest for each manuscript is the average score for the two blinded referees versus the average score for the unblinded ones. Also of interest is whether the variability between blinded reviewers is similar to the variability between unblinded ones.

## 5. AUTHOR CHARACTERISTICS

Proponents of blind refereeing are concerned, in part, with the possibility of biases in the review process based on author gender, nationality and professional stature. For example, it might be that the unblinded referees would be more likely to rank highly a manuscript from a well-known author or from a highly regarded institution. Conversely, one might anticipate that blinded reviewers might give a higher ranking than unblinded reviewers to a manuscript from a new graduate or from a lesser known institution. In particular, while the null hypothesis is that there is no difference in rating between blinded and unblinded reviewers, the alternative, both in direction and magnitude, may depend on the spectrum of manuscripts under review. To obtain a baseline assessment of author and institution reputation, we propose that at the time of receipt of the manuscript, the Editor rank the "stature" of the authors and institutions on four individual scales, each scored from 1 (well known and highly regarded) to 5 (unknown, or not well regarded). The four scales would be (1) stature of first author, (2) stature of best-known author, (3) stature of institution of first author and (4) stature of best-known institution. Gender and country of origin of the first author will also be recorded by the Editor.

To assess referee perception about the authors, we suggest that unblinded referees submit their guesses as to the gender and country of origin of the first author, and that blinded referees submit a guess as to authorship of the paper as well.

## 6. THE PILOT STUDY

The main charge of this design committee is to propose a pilot study of blinded refereeing. The main goals of the pilot study are (1) to assess the feasibility of conducting a full-scale study of the use of blind refereeing for the journals and (2) to estimate the magnitude of inter-rater variability within the two arms of the study and variability of scoring between reviewers in

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.*