NEGLECTED CONSIDERATIONS IN THE ANALYSIS OF AGREEMENT AMONG JOURNAL REFEREES

L. L. HARGENS,* J. R. HERTING**

*Department of Sociology, University of Illinois,
702 South Wright Street, Urbana, IL 61801 (USA)
**Department of Sociology Stanford University,
Stanford, CA 94305 (USA)

(Received September 4, 1989)

Studies of representative samples of submissions to scientific journals show statistically significant associations between referees' recommendations. These associations are moderately large given the multidimensional and unstable character of scientists' evaluations of papers, and composites of referees' recommendations can significantly aid editors in selecting manuscripts for publication, especially when there is great variability in the quality of submissions and acceptance rates are low. Assessments of the value of peer-review procedures in journal manuscript evaluation should take into account features of the entire scholarly communications system present in a field.

Introduction

Referees' assessments of manuscripts submitted to scholarly journals play a key role in determining editors' eventual decisions about the fates of those manuscripts. For example, in a study of the determinants of the final dispositions of manuscripts submitted to the American Sociological Review, Bakanic et al. found that the zero-order correlation between the mean of referees' overall recommendations and final manuscript dispositions equalled 0.81.1 When they regressed final manuscript dispositions on mean referees' recommendation and 23 other variables describing features of manuscripts and the manuscript-evaluation process, the standardized partial regression coefficient for the mean referees' recommendation was only slightly smaller, 0.74, than its zero-order correlation.2 Thus, it is not surprising that both researchers and commentators have devoted considerable attention to the topic of inter-referee agreement on assessments of manuscripts.3-7

This literature, and associated work on the selection of papers for presentation at scholarly meetings8-9 and the selection of research proposals for funding10-11 has four characteristics. First, it is fragmented. Although psychologists tend to be more active researchers in this area than scholars from other disciplines, studies on the
topic have been reported by researchers in a wide range of fields. Much of the research on journal referee agreement has been carried out by journal editors, who have privileged access to data, and has been reported in their annual reports or in special editorials. Few researchers have published more than one report on the topic, and the literature largely consists of opportunistic uses of available data rather than sustained efforts to explore the topic. Second, the literature exhibits a high ratio of commentary to data. A recent paper on journal referee agreement revealed that as of 1983 complete data for a crosstabulation of referees' assessments had been published for only two journals! Third, nearly all of this research takes a "psychometric" perspective on the topic. This perspective views submitted manuscripts as varying in their "merit" or "publishability," and asks whether referees' assessments can be expected to provide editors with reliable information about this latent variation. To the extent that referees' assessments index manuscripts' relative merit, referees' independent evaluations of manuscripts should be correlated.

Finally, it has been the fashion, in recent years at least, to emphasize, if not exaggerate, the low level of agreement between peer-reviewers' evaluations. In this paper we discuss several points that researchers should consider in carrying out and interpreting studies of referee agreement. Some of these points are general methodological principles; for example, that variables with restricted variation tend to have low correlations with other variables, and that measures with relatively low validities can still be quite useful in selection procedures. Other points noted below involve structural and statistical characteristics of the manuscript evaluation process itself; such as how editors' practices in choosing referees may limit the observed level of agreement between referees' evaluations, and how the degree of literature concentration in a field may counteract the effects of referee unreliability at an individual journal. Those who report or comment on referee agreement studies typically neglect one or more of the points we discuss. In general, we believe that taking them into account leads one to be more sanguine than many recent discussions of the value of peer-review procedures in evaluating manuscripts.

Is referee agreement low?

We note first that many studies of referee agreement were predestined to yield low levels of agreement because they were based on samples of papers with restricted variation in merit. For example, in an early study of agreement between expert judges' evaluations of papers, Bowen et al. took advantage of data yielded by a competition for "best paper" among papers submitted to the Division of Consumer Psychology for presentation at the 1971 convention of the American Psychological