Research Notes

Ratings and Rankings: Multiple Comparisons of Mean Ratings

William E. McGrath

Ranking of journals or other objects according to mean ratings computed from an opinion survey is shown to be inappropriate if a test of significance shows no difference between them. A Scheffé test for comparisons of mean ratings of journals ranked by Kohl and Davis [C&RL 46:40–47 (Jan. 1985)] *was performed. The results indicate no significant difference between means. Confidence intervals for every adjacent pair of journals in the list of ratings by ARL directors were also computed. The results indicate that every adjacent interval overlaps, and that the means are essentially tie scores. Treating them as significantly different, therefore, is a Type 1 error.*

Rank ordering of mean ratings, a common practice in library science research, can lead to serious Type 1 errors if the mean ratings are not first submitted to tests of significance. "Type 1" errors are those in which a hypothesis assuming no difference between two means, say, is actually true but is treated as untrue by the researcher. In turn, Type 1 errors, if not recognized, may lead to unjustified social or administrative actions or other errors of judgment or policy.

Two examples will illustrate. The first is from my own research some years ago, which inconclusively attempts to correlate mean ratings of subject-area characteristics (computed from a 10-point scale) with variables of library circulation.[1] The absence of strong correlations may be attributed to the probable absence of significant differences between the mean ratings of subject areas. Had those differences been tested, the limitations of my design might have been realized. Fortunately the long-term consequences were as negligible as the correlation, as I had merely failed to build good theory.

The second example appears in an article by Kohl and Davis.[2] These authors asked ARL library directors and deans of accredited library schools to rate thirty-one library journals in terms of their importance to evaluations of publications by librarians or faculty being considered for promotion and tenure. Each journal title was rated by each respondent on a 5-point Likert scale. The authors computed the mean rating of each journal, then ranked the journals according to these means. As in my own research, the authors did not test to determine whether means were significantly different from each other—although they did compare directors' ratings to deans' ratings. Without such a test there is no evidence that one mean rating is any different from any other.

The rankings in question appear in their

William E. McGrath is Associate Professor in the School of Information and Library Studies, State University of New York at Buffalo, New York 14260.

table 1.[3] These ranks *seem to assume* that each mean is different—for example, that the mean for *Library Quarterly*, 4.4048, is different from that for *Journal of Academic Librarianship*, 4.3810—when in fact they are probably not different. That is precisely the same error cited in the first example—a Type 1 error.

Kohl and Davis, however, did seek to avoid Type 1 errors, first by performing t-tests for the differences between the means of ARL directors and library school deans, then by looking at internal consensus. They report the results of that test in their table 2. They conclude, that because deans and directors appear to agree on their ratings of journal "importance," there is a "perceived hierarchy of journal prestige."

However, their Type 1 errors are between *journals*, not between deans and directors. Thus, their finding of a "perceived hierarchy of journal prestige" is not supported. Although a perceived hierarchy may exist, it cannot be determined from their table 1. Therefore, acceptance of these journal ranks at face value for the purpose of determining promotion and tenure of librarians and faculty could lead to inappropriate evaluation.

The small visual differences between the means in table 1 and the small sample size from which the journal means were computed also cast suspicion on conclusions drawn from them. The Scheffé test is appropriate for all possible comparisons.[4] The data reported in their tables 1 (mean ratings) and 2 (sample sizes and standard deviations) make it possible to compute an overall *mean square within (MSw)*, which is required to compute an *F* statistic, which, in turn, is required to perform the test. The equation for *F* is

$$F = \frac{(M_1 - M_2)^2}{MSw \left(\frac{1}{n_1} + \frac{1}{n_2}\right)(k-1)}, \qquad \text{(a)}$$

with $df = k-1, N-k$.

Working backwards, it is possible to compute *MSw* from the statistics reported in table 3, as follows:

$$MSw = (\Sigma S_j^2 n_j - \Sigma S^2_j)/(N-k), \qquad \text{(b)}$$

where S^2_j and n_j are the squares of the standard deviations and the sample sizes for each journal respectively. A sample size of 42 for each journal, reported in Kohl and Davis' table 3, is assumed in computing the above equations.

The Scheffé test was performed on means of ARL directors' ratings (left column of table 1) but only for the journals in Kohl and Davis' table 3, which contains the standard deviations necessary for the computation. From (b) above, $MSw = 2.23$. This value was used in (a) to compute *F* values for the Scheffé tests appearing in table A.

For no adjacent pair of journals did the computed values of *F* exceed the test value of 1.57, indicating true null hypotheses in every comparison—i.e., that the means for every adjacent pair in the list are not significantly different from each other. Not until the journal at the top of the rankings, *College & Research Libraries*, was compared with one well down in the list, namely *Library and Information Science Research*, was a significant difference observed. Furthermore, *Library and Information Science Research* is not significantly different from the journals following it in the list. This general lack of significance does not appear to support the rationale for strict ranking of these journals. At best, one might postulate two clusters of journals, with each journal in the first cluster essentially tied for first place and each in the second cluster tied for second place. To paraphrase *Consumer Reports*, journals within clusters are approximately equal in importance.

Nearly identical results were obtained when a t-test for independent samples (though these samples may not be truly independent) was performed, again working backward from the standard deviations to obtain sums of squares and standard errors of the differences between each pair of means.

Finally, confidence intervals for all means in the ARL directors' list were computed, again at the .05 significance level. For every journal, the confidence interval overlapped the one above it and below it. For example, the lower and upper limits for *C&RL* were 4.60 and 4.87, respectively, while the lower and upper limits for *LQ*

TABLE A

SCHEFFÉ TEST FOR DIFFERENCES BETWEEN PAIRS AND CLUSTERS OF JOURNAL MEANS

Journal Title	Mean	Pair-wise F value*	Possible Clusters†
Coll. & Res. Libr.	4.7381	0.06	Possible Cluster 1
Libr. Quart.	4.4048	0.00	
J. Acad. Libr.	4.3810	0.00	
Libr. Res. & Tech. Serv.	4.3810	0.01	
Library Trends	4.2381	0.00	
Info. Tech. and Libr.	4.1429	0.00	
JASIS	4.0952	0.03	
Library Journal	3.8571	0.06	
American Libraries	3.5000	0.01	
RQ	3.3810	0.02	
Special Libraries	3.1667	0.04	.... Possible Cluster 2
Libr. & Info. Sci. Res.	2.8810	0.06	
Collect. Management	2.5238	0.18	
Info. Proc. & Mgmnt.	1.9286	0.02	
School Library Journal	1.7381	0.01	
Intern. Libr. Rev.	1.5714	0.00	
Micrographics Today	1.5714	0.00	
School Library Media Q	1.5714	0.00	
Intern. J. Law Libraries	1.5476	0.00	
Law Library Journal	1.5238	xxxx	

*F(df: k − 1 = 19, N − k = 820), .05 level = 1.57.

The F value refers to pairs of titles: the title listed and the one immediately following. Thus, the first F listed, 0.06, refers to *College and Research Libraries* and *Library Quarterly*. F values must exceed 1.57 to be significant. None are.

†Means for journals within "possible clusters" are not significantly different from each other. But the first title in cluster 1 (*C&RL*) is significantly different [*F*(.05) = 1.71] from the first title in cluster 2 (*Library and Information Science Research*), while the last title in cluster 1 (*Special Libraries*) is significantly different (F = 1.71) from the last title in cluster 2 (*Law Library Journal*), clusters 1 and 2 overlap each other with *Special Libraries*. The difference between the average of cluster 1 and the average of cluster 2 is significant [F(.05) = 22.7].

were 4.09 and 4.72. Clearly, the upper limit of *LQ* falls well within the interval for *C&RL*, indicating that their means cannot be distinguished from each other.

Visual inspection of the means for library school deans' rankings (right column of table 1) suggests that few significant differences would be found between adjacent journals in that list either.

This analysis suggests that ranking average ratings without submitting them to appropriate tests of significance cannot be trusted. Such tests are necessary even when data are trustworthy—for example, when the sample is large, or when it otherwise represents the population with a high degree of confidence. Here, a distinction should be made between performing tests of significance to guard against *sampling* errors on the one hand and *measurement* errors on the other. Here, the rating scores can properly be considered as measurements subject to error. For example, an average score can hide a great diversity of opinion. If we ask 100 respondents to rate journals on a 1-to-5 scale, a particular journal could receive an average of 3.0 in several ways. At the extremes, all respondents could give the journal a rating of 3; or 50 respondents could give a rating of 1; and 50, a rating of 5. Both scenarios produce an average of 3.0, but the first represents exact consensus. In the second, the average score hides a considerable degree of measurement error. In fact, in the second scenario *no* individual respondent gives the journal a rating of 3.0 and we might well question whether a real consensus exists that a journal with a rating of 3.0 is really higher than one with a rating of 2.9.

Kohl and Davis sprinkle cautions throughout their study, noting that it has "important limitations" that must be considered "to maintain a proper perspective on the findings." Perhaps the major caution should address the use of these or similar ranks for determining tenure and promotion.

If journal prestige and importance must be studied, then many related questions—including those raised here and by Kohl and Davis—must also be studied. Which journals do the larger population of non-

ARL directors and ACRL members feel are important? What is the relationship between a respondent's own specialized area and the subject area of the journal being rated? What are the correlates of "prestige" or "importance"? Can prestige or importance be predicted from other variables? What is the basis for equating prestige and importance? Is prestige a variable of real utility, or does it merely make an author feel good? Do studies of prestige contribute to the knowledge base of our profession? Or does the knowledge base contribute to prestige? Prestige is not a guarantee of quality, say Kohl and Davis. Likewise quality is not a guarantee of prestige. Then what is quality, and what is the relationship between prestige and quality? Kohl and Davis suggest citation analysis; other kinds of impact should also be examined. It seems that whenever we attempt to measure attitudinal variables, we can never really pin them down without reference to behavioral variables. Understanding of behavioral variables has much the greater potential for contributing to good theory.

In conclusion, whenever rating scores are used to produce rankings of items being rated, those rankings should be subjected to appropriate tests of statistical significance.

REFERENCES AND NOTES

1. William E. McGrath, "Predicting Book Circulation by Subject in a University Library," *Collection Management* 1, no.3/4:7–26 (Fall/Winter 1976–77). Average ratings in this research were for the variables *Hard/Soft, Pure/Applied,* and *Life/Nonlife.*
2. David P. Kohl and Charles H. Davis, "Ratings of Journals by ARL Library Directors and Deans of Library and Information Science Schools," *C&RL* 46:40–47 (Jan. 1985).
3. All references to tables are to Kohl and Davis except for table A.
4. John T. Roscoe, *Fundamental Research Statistics for the Behavioral Sciences* (New York: Holt, 1975), p.313.

Authors' Reply

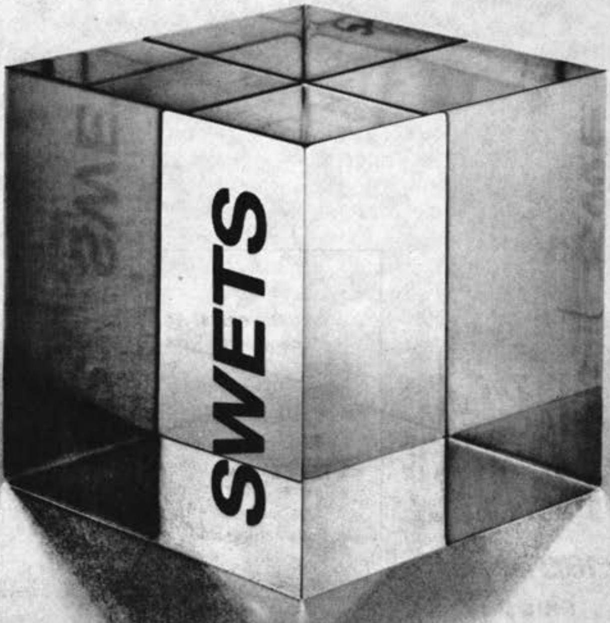
David F. Kohl and Charles H. Davis

We read William McGrath's comments on our study with considerable interest. Our only concern is that in order to make his point he has to make us say more than we were, in fact, comfortable saying. It frankly never occurred to us that anyone would take the listing in Table 1 as some kind of precise ranking where "each mean is different," since that is obviously not the case. Not only did a number of the journals listed in Table 1 have identical means and were, in those cases, "ranked" in alphabetical order but in addition we present two other possible "rankings" which vary in detail from the lists in Table 1. The point of the article, which was fairly explicitly made, was not that any one journal stood in a specific relationship to any other journal, but that a clearly recognizable general pattern did exist with some journals consistently emerging toward the top, others toward the middle, and others toward the bottom.

David F. Kohl is Assistant Director for Public Services, University of Colorado, Boulder, Colorado 80309. Charles H. Davis is Professor, Graduate School of Library and Information Science, University of Illinois, Urbana, Illinois 61801.

In fact, Professor McGrath's own analysis seems to confirm this general hierarchy or, as he calls it, clustering. It should be noted that he finds this very general clustering (into two groups) using the Scheffe test—the most conservative test of this kind possible. A less restrictive test such as the Duncan, Tukey, etc., would invariably have suggested finer distinctions among the journals. The issue, which McGrath's comments may obscure, is not whether there is or is not some hierarchy or ranked clustering but how fine the gradations of the hierarchy or clustering are.

We agree with McGrath's point that averages don't necessarily constitute a detailed ranking and hope that his comments may help prevent a misreading of Table 1 of our study by casual readers. We do feel, however, that his misinterpretation of Table 1 created a bit of a straw man in our case.



SWETS