Journal of Educational Measurement
Spring 1987, Vol. 24, No. 1, pp. 72–76

Note About Possible Bias Resulting When Under-Statisticized Studies are Excluded from Meta-Analyses

Julian C. Stanley
Johns Hopkins University

Reviews and meta-analyses of research on a given topic may exclude a sizable percentage of reports because they do not lend themselves to the type of summarizing procedures used. If the excluded articles contain relevant information, this may bias the conclusions of the analysis. It seems likely that, when computing statistics from their data, researchers will need to consider this aspect. A simple illustration of how that can sometimes be done readily is presented. A robust correlation coefficient easily computable from published data is shown to indicate a sizable relationship that is contrary to the main conclusion of a meta-analysis.

Authors of original research should be aware that their studies will tend to be excluded from meta-analyses if their results are not presented in a statistical form amenable to that type of analysis. These authors should consider that sometimes data can be reconfigured readily to produce such statistics. Authors of meta-analyses should consider that exclusion of "under-statisticized" reports can bias their results (e.g., see the "file drawer" problem, Hedges & Olkin, 1985), and that such reconfiguration may be possible with published data. The following simple example illustrates such a reconfiguration and provides information about its robustness. It also illustrates that the maximum effect possible for a given data set may be constrained greatly by the differential shapes of the two score distributions.

Example

Stanley (1976) showed that the mathematical section of the College Board's Scholastic Aptitude Test (SAT-M) predicted the scores of 11th-graders in a mathematics contest 2 or 3 years later far better than the students' mathematics teachers did. Fifty-one students participated in the contest, which was conducted by a university mathematics department. The 60 SAT-M items were all five-option multiple-choice measures of mathematical reasoning ability, with content not beyond that usually taught in the ninth grade but stressing novelty rather than routine application of learned rules. As its title indicates, SAT-M is supposed to test "aptitude" more than "achievement," a much-debated distinc-

I thank Camilla P. Benbow, Linda E. Brody, Robert A. Gordon, Lois S. Sandhofer, Barbara S. K. Stanley, Wendy M. Yen, and an anonymous *JEM* referee for statistical and editorial assistance. This work was supported by the U.S. Department of Education.

tion (e.g., Green, 1974). Questions in the mathematics contest were open-ended problem sets of the "clever" kind usually devised by college mathematicians.

Of the 51 participants, 10 were those who competed from among the 17 who, almost at the last minute, were invited to enter the contest by the Study of Mathematically Precocious Youth (SMPY) solely because 2 or 3 years earlier they had scored well (typically, in the 600s as eighth-graders) on the SAT-M in one of SMPY's annual talent searches, and now were probably 11th-graders. Three of these 10 happened also to have been nominated by their teachers. None of the 10 ranked lower than 23.5 out of the 51 in the contest. The number 1 scorer in the mathematics contest had been nominated only by SMPY, the number 2 scorer by both SMPY and his teacher, the number 3 scorer only by SMPY, and so forth. More complete information is contained in Stanley's 1976 work.

Recently, an unpublished review (Hoge & Cudmore, n.d.) of the validity of

Table 1
Agreement of Mathematics Contest Rank with Nomination Score

Contest Rank (1 = Highest)	Nomination Score (1, 0.5, 0)	Best-Possible-Agreement Score
1	1	1
2	0.5	1
3	1	1
4	0	1
5.5	0	1
5.5	0.5	1
7	1	1
8	1	1
9	0	1
10	0	1
11	0	0
12	1	0
13	0	0
14	0	0
15	0	0
16.5	0	0
16.5	1	0
18	0	0
19	0.5	0
20	0	0
21	0	0
22	0	0
23.5	0	0
23.5	1	0
25 to 51	0	0

Note. r between contest rank and score is -.52 for the (1, 0.5, 0) score and -.69 for the best-possible-agreement score.

judgments of giftedness made by teachers versus those made by tests cited Stanley (1976) but did not use its results because they were not expressed as a statistic such as r, t, or F. In the following section a nominating variable is introduced that allows calculation of coefficients of correlation of the ranks of scores from the mathematics contest with the nominating process. This nominating variable, or a similar one, could have been created by the authors of the meta-analysis from the information in the original article.

Nominating Variable

A nominating variable was devised as follows: 1 if the student was nominated only by SMPY, 0.5 if nominated by both SMPY and a teacher, and 0 if nominated only by a teacher. Represented by the notation (1, 0.5, 0), it has intrinsic order with respect to the degree that SMPY nomination was involved (exclusively, equally, and not at all, respectively). Scores based on this nominating variable and ranks from the contest results are presented in Table 1. The Pearson product-moment r proves to be $-.52$, indicating that with trichotomization there is a fairly strong tendency for SMPY-nominated contestants to rank higher in the math-department contest than do contestants nominated by their teachers (see Table 1).

Examination of alternative nominating variables indicates that the correlation between nomination and contest rank is quite robust. For example, dichotomizing as 1 = nominated by SMPY versus 0 = nominated only by a teacher (1, 0) produces an r of $-.54$. Trichotomizing as 0.5 = nominated only by SMPY, 1 = nominated by both, and 0 = nominated only by teacher results in an r of $-.51$.

The equal-spacing scale of the (1, 0.5, 0) variable is arbitrary. However, most other scaling would change the r little. For example, using the Kelley (1947, p. 297, Formula 8:27) normalizing transformation of the three ordered categories to get the z-score deviation of the mean of the category from the mean of the unit normal distribution (which is 0) yields z-scores of 1.60 for sole nomination by SMPY, 0.97 for joint nomination, and -0.34 for teacher-only nomination.[1] The

[1]Most readers may not have ready access to Kelley's book or be familiar with this technique, which long ago formed the basis for the now much misunderstood "grading on the normal curve." It involves first changing the ordered categorical frequencies into proportions. Here they are 41/51 = 0.8039 for 0, 3/51 = 0.0588 for 0.5, and 7/51 = 0.1373 for 1. The mean z-score for each category is then calculated, assuming an underlying normal distribution. Calculating these is equivalent to obtaining the means of truncated normal variables (e.g., see Johnson & Kotz, 1970, p. 81.) In this calculation the *heights* of the two ordinates bounding the left and right side of each category are obtained from a table (e.g., Kelley, 1938) or calculated $[(1/\sqrt{2\pi})\exp(-z^2/2)]$. For the 0 category these heights are 0 and 0.2767. For the 0.5 category they are 0.2767 and 0.2196. For the 1 category they are 0.2196 and 0. To get the normal-distribution z-score for the mean of the 0 category, compute $(0 - 0.2767)/0.8039 = -0.34$. For the 0.5 category, compute $(0.2767 - 0.2196)/0.0588 = 0.97$. For the 1 category, compute $(0.2196 - 0)/0.1373 = 1.60$. Where formerly one had only ordered categories with their respective percentages there are now scores.

score for the first nomination category differs from the second by 0.63, but the second differs from the third by 1.31, more than twice as much. If one uses these *z*-score values (1.60, .97, −.34) rather than the (1, 0.5, 0) scale, the *r* changes from −.52 to −.53. The two sets of nomination "scores" correlate .994.

Even the intraclass coefficient of correlation r_i (Stanley, 1971, p. 426), which treats the categories as unordered, produces a similar result. For these data it is .52. Thus, despite the crude appearance of the nominating "variable," the various *r*s are remarkably similar. This may occur largely because only 3 of the 51 cases are in the nominated-by-both category.

Importance of Score Distributions

A usual way to interpret a Pearsonian *r* is to square it (i.e., yielding .26–.29 here) and consider that as the proportion of variance in the outcome scores that is accounted for by a linear relationship. More-than-25% predictability on the basis of a 75-minute SAT-M administered several years earlier is impressive.

These percentages do not, however, fully reveal how high the relationship actually is, because the maximum *r* that could possibly have occurred is only −.69. This maximum occurs if the 10 students who had been nominated only by SMPY had ranked 1, 2, . . ., 10 in the contest (see Table 1). Thus, the squared maximum *r* is .48, not the 1.00 that would be *possible* if the two sets of scores had the same shape. It is well known that great attenuation of the maximum possible value of an *r* occurs when the shape of one distribution differs considerably from that of another (e.g., see Carroll, 1961, and Ozer, 1985).

Conclusion

Quantifying one's results more fully than Stanley (1976) did is important for informing readers about their magnitude, and has become necessary if one's findings are to become part of quantitatively oriented reviews or meta-analyses (e.g., see Glass, McGaw, & Smith, 1981). If such quantification does not occur, research may fail to have the impact it deserves; omitting such quantification may also bias the conclusions of the meta-analysis. At the very least, the original data should be retained for several years by the author(s) of an article and made available to any meta-analyst who requests them.

References

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika, 26,* 347–372.

Glass, G. V, McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Green, D. R. (Ed.). (1974). *The aptitude-achievement distinction.* New York: McGraw-Hill.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Hoge, R. D., & Cudmore, L. (No date, ca. 1984). *The use of teacher-judgment measures in the identification of gifted pupils.* (Department of Psychology Research Rep., not numbered). Ottawa, Ontario: Carleton University.

Johnson, N. L., & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions* (Vol. 1). New York: Wiley.

Kelley, T. L. (1938). *The Kelley statistical tables.* New York: Macmillan.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge, MA: Harvard University Press.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97,* 307–315.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356–442). Washington, DC: American Council on Education.

Stanley, J. C. (1976). Test better finder of great math talent than teachers are. *American Psychologist, 31,* 313–314.

Author

JULIAN C. STANLEY, Professor, Department of Psychology, and Director of the Study of Mathematically Precocious Youth at Johns Hopkins University, Baltimore, MD 21218. *Degrees:* BS, Georgia Southern College; EdM, EdD, Harvard University. *Specializations:* mathematical precocity, educational acceleration, individual differences, test theory.