ED 170 353 TM 008 775

AUTHOR Lai, Morris K.

TITLE Rules of Thumb from the Literature on Research and Evaluation.

PUB DATE Apr 79

NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Evaluation Methods; *Research Criteria; *Research Methodology; *Research Problems; *Testing

ABSTRACT

Practical advice on frequently asked questions dealing with research and evaluation methodology is presented as rules of thumb, with citations to the author's sources. A statement in the literature is considered a rule of thumb if it meets one of the following criteria: (1) it is specifically called a rule of thumb; (2) it contains numbers in place of algebraic symbols; or (3) it contains a reference to previous successes using a particular level. The rules included here deal with article title, budgeting for research staff, test difficulty and discrimination, distribution, significance, Fisher test, reliability of gain scores, interrater reliability, item construction, testing time, response rate, observation, sample size, sampling, skewness, test wiseness, test-retest reliability, and test revision. (MH)

Rules of Thumb from the Literature on Research and Evaluation

MORRIS K. LAI
University of Hawai'i
1776 University Ave.
Honolulu, HI 96822

Paper presented at the annual meeting of
the American Educational Research Association,
San Francisco, April, 1979
Session 13.03

Rules of Thumb from the Literature on Research and Evaluation

MORRIS K. LAI
University of Hawai'i

Introduction

Almost every researcher and evaluator has had the frustrating experience of seeking quickly needed practical advice, but ending up with either long drawn out discourses or complicated formulas that are of little practical use. How many times have the following questions been asked, but no usable answers have been given: How large a sample should I use? How many items should be put on the test? How reliable should the test be? What constitutes an educationally significant difference? Even if ballpark answers are requested to questions like these, more often than not, practical answers are not forthcoming. Sometimes when answers are suggested, it is evident that personal bias has in a large way influenced the response.

Although consultants or textbooks may validly respond to a practitioner's questions by saying "It depends," oftentimes a reasonable estimate or ballpark figure would be more appropriate. In fact, a rule of thumb based on previous empirical and theoretical results may in many cases be even more "correct" than a rule that results from exact, complicated formulas which are based on questionable assumptions.

The following question is perhaps one of the best ways of illustrating the perspective being taken in this paper: If you were developing a test and wanted to know how many subjects should be in a formal tryout for item analyses, which type of response would you prefer? a) "It all depends, but here is a 600 page book on tests and measurements," or b) "Henrysson (in Thorndike's _Educational measurement_, 1971) recommends that at least 300 subjects be used." Although some will insist that the first response is more defensible, many others would benefit much more from the second response.

Method and data source

The uncovering of rules of thumb is of course a never-ending task; however, as a start, the author went through as many educational research and evaluation books that he could find in a) his own collection, b) in the University of Hawai'i library, c) in the collections of colleagues, and d) in the many research and evaluation projects at the Curriculum Research and Development Group of the University of Hawai'i. Concurrently a search of the most appropriate journals (e.g., Psychological Bulletin, Review of Educational Research, American Educational Research Journal) was carried out. For the past five years all promising AERA meeting articles were sent for and read. Finally a retroactive ERIC search was carried out. In order to make the task of reasonable one, publications before 1970 were in general not included.

Given this vast amount of data, it was necessary to skim rapidly over all parts of the material which did not constitute rules of thumb. In selecting rules of thumb the following definition was used (admittedly with some flexibility)-- a statement was considered a rule of thumb if any of the following were true: a) it was identified specifically as a rule of thumb, b) it was a suggestion that contained actual numbers in place of algebraic symbols (e.g., "p should be between .2 and .8"), or c) a reference was made to previous successes using a particular level (e.g., "So and-so found that at least 1000 subjects were needed for a national sample.").

Among the statements that were not classified as rules of thumb were 1) results based on a single study (and reported as such), 2) general recommendations (e.g., "Involve the evaluator at the beginning of the project.") 3) rules which were likely to be or become outdated (e.g., "Expect to spend $________ in carrying out the following task.) 4) rules whose content was too exotic or unique to be of interest to many practitioners (e.g., Glass et. al., 1975, page 96-- On partial autocorrelations: "Perhaps only the first two or three autocorrelations can be adequately estimated from (5.36) with even relatively long series (n=50 to 100).")

Those statements which passed the definition screening were then recorded on 4x6 cards and classified by author as well as content area. Where conflicting rules of thumbs were found, all were included. In many cases authors presented rules which were referenced to other authors. A decision was made to cite both the reference in which the rule was found as well as the author to whom the rule was attributed; however, for pragmatic reasons, only the reference in which the rule was actually read will be listed at the end of the complete treatise. (e.g.Ebel in Ahmann & Glock, 1971: The difficulty level of test items should be between .40 and .70)

In doing the research it became apparent that in the best tradition of oral transmission of culture many rules of thumb have been passed down through the years, oftentimes without reference to the rule's originator(s). When these cases have arisen no serious attempt has been made to track down the true source. Instead an often arbitrary representative has been selected to receive credit or blame, if not as the originator of the given rule, then as a perpetuator.

In my attempts to develop and describe the methods used in compiling a collection of rules of thumb I have been somewhat influenced by Jackson (1978) who forcefully argued that methods used in reviewing research should be made explicit. He also discussed the many ways in which the methodology of such reviews could be improved. In the current attempts to put together a compendium of rules of thumb, it became quite apparent that the methodology of compiling was perhaps as important as the rules themselves. As alluded to earlier, the method used could affect the number and type of rules selected, the classification of the rules, the author referencting, etc.

The lengthiness of the list of rules of thumb together with their references precludes a complete presentation in this paper. Instead the following examples are given to help the reader decide whether or not to obtain the comprehensive compendium that will be available in the near future.

Article title. Maximum length should be 12-15 words (APA, 1974, p. 14).

Budget. For RFP's (requests for proposal), person years are translated at times into $25,000 to over $50,000 (Scriven & Roth, 1977, p. 17).

Difficulty level. The difficulty level for items on classroom tests should be between .4 and .7 (Ebel, R. L. in Ahmann & Glock, 1971).

Discrimination index. A reasonably good achievement test item should have an index of at least .30 (Ahmann & Glock, 1971, p. 189).

Distribution. Group observed outcomes into 8 to 12 equal-width intervals (Marascuilo, 1971, p. 179).

Educationally significant. A difference is educationally meaningful if it is ≥ 1/3 of a standard deviation (or sometimes ≥ 1/4 s.d.) or rate of growth produces a post percentile greater than the pre percentile by one standard error (Tallmadge, 1977, p. 34).

Fisher Test. If N < 20, use the Fisher Test (instead of X^2) in all cases (Siegel, 1956, p. 110).

Gain score mean reliability. For N ≥ 30, gain score means are probably quite reliable (Martuza, 1977, p. 141).

Interrater reliability. a) Should be at least .70 (Borg & Gall, 1971, p. 235). b) Observers should be in perfect agreement 80% of the time (Borich, 1974, p. 259).

Item construction. Construct 20% more items than are needed (Aiken, 1976, p. 30).

Item test time. Average high school student should be able to answer two true-false items, one multiple choice item, or one short answer item per minute of testing time (Gronlund, 1971, p. 240).

Non-respondents. A questionnaire non-respondent rate of less than 20% can be reasonably ignored (Isaac & Michael, 1971, p. 43).

Observations. For practical purposes of estimation two repeated independent observations on any person are typically sufficient (Novick & Jackson, 1974, p. 86).

Response rate. For mail surveys, an 80% return is acceptable (Sudman, 1976, p. 30.)

Sample size-cohort study. Need an N of 500-1000 per cohort for a 3-year study (Cooley & Lohnes, 1976, p. 137).

Sampling correction. Finite population correction can be ignored whenever the sampling fraction does not exceed 5% (and for many purposes even if it is as high as 10%) (Cochran, 1977, p. 25).

Skewness, computation of. Seldom advisable to compute g_1 (measure of skewness) when $N < 100$ (McNemar, 1969, p. 27).

Test practice effect. Usually improves scores at the second testing by no more than $\frac{\sigma}{5}$ (Anderson, et. al., 1975).

Test-retest reliability. Should be $>.85$ (Massad, 1977, p. 243).

Test revision. Given a fairly broad level of ability, 5 to 8 students can give considerable revision help (Bloom, et. al., 1971).

Some precautions are in order for users of rules of thumb. In many circumstances it will be important to get more details to insure that an appropriate rule of thumb is being applied. Some rules of thumb may represent remnants from mythology or traditional ignorance (cf. Aristotle). Other rules of thumb may have come from persons brash enough to generalize from a single study. A given statement might also have hidden in it the peculiar values of an individual (eg, a relatively conservative writer may present more stringent rules of thumb). The user of the rule must, therefore, take into account the source or perpetuator of the rule.

Despite these cautions it appears that rules of thumb can be of substantial benefit, especially to the practitioner. Lest I incur the wrath of the union of high-cost consultants, I would also add that personal expert advice can still be extremely valuable. In fact it wouldn't be a bad idea to approach research/evaluation problems through the use of a combination of a consultant and a rule of thumb. At least with this combination, the researcher will never end up with the far-too-frequent occurrence of paying for advice, but not knowing what to do next. Of course the practitioner should make sure that the consultant doesn't end up charging a lot for merely supplying a rule of thumb that the user already knows about.

In keeping with the spirit of this paper, I will end with two final rules of thumb: 1) 95% of paper presenters at conferences go over their allotted time (or should it read: Not enough time is allotted for 95% of the paper presentations), and 2) Listeners or readers start to fall asleep after about 8 pages of a conference paper (Lai, 1979 AERA Conference). In order to retain my status as a bonafide user of rules of thumb, I hereby end this presentation and invite all of you to join the society of thumb rulers.

REFERENCE LIST

- Ahmann, J. S., & Glock, M. D. Evaluating pupil growth. Boston: Allyn & Bacon, 1971.

- Aiken, L. R. Psychological testing and assessment. Boston: Allyn & Bacon, 1976.

- American Psychological Association. Publication manual. Washington, D.C.: American Psychological Association, 1974.

- Anderson, S. B., Ball, S., Murphy, R. T., & Associates. Encyclopedia of educational evaluation. San Francisco: Jossey-Bass, 1975.

- Bloom, B.S., Hastings, J. T., & Madaus, G. F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

- Borg, W. R., & Gall, M. D. Educational research. New York: David McKay, 1971.

- Borich, G. D. (Ed.). Evaluating educational programs and products. Englewood Cliffs, NJ: Educational Technology, 1974.

- Cochran, W. G. Sampling techniques. New York: John Wiley & Sons, 1977.

- Cooley, W. W., & Lohnes, P. R. Evaluation research in education. New York: Irvington, 1976.

- Gronlund, N. E. Measurement and evaluation in teaching. New York: Macmillan, 1971.

- Henrysson, S. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

- Isaac, S., & Michael, W. B. Handbook in research and evaluation: San Diego: Robert R. Knapp, 1971.

- Jackson, G. B. Methods for reviewing and integrating research in the social sciences. Washington, D.C.: George Washington Unviersity, 1978.

- Marascuilo, L. A. Statistical methods for behavioral research. New York: McGraw-Hill, 1971.

- Martuza, V R. Applying norm-referenced and criterion-referenced measurement in education. Boston: Allyn & Bacon, 1977.

- Massad, C. E. (Ed.). Resource notebook of information for assessment and evaluation processes for T & E. Princeton: Educational Testing Service, 1977.

- McNemar, Q. Psychological statistics. New York: John Wiley & Sons, 1969.

- Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974.

- Scriven, M., & Roth, J. Evaluation thesaurus. Pt. Reyes, California: Edgepress 1977.

[illegible] for the behavioral sciences. [illegible]

[illegible] sampling. New York: Academic Press, [illegible]

[illegible] The Joint Dissemination Review Panel IDEABOOK. [illegible] [illegible] Printing Office, 1977.