P9

Analysis of Attitudinal Data:
Dealing with"Response Error"

Janet McCollum

Houston Independent School District

Bruce Thompson

University of New Orleans

McCollum, J., & Thompson, B. (1980, April). *Analysis of attitudinal data: Dealing with "response bias"*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Abstract

Response error refers to the tendency to respond to items based on the perceived social desirability or undesirability of given responses. Response error can be particularly problematic when all or most of the items on a measure are extremely attractive or unattractive. The present paper proposes a method of (a) distinguishing among preferences even when all items are uniformly attractive, and (b) computing *intraindividual* reliability coefficients to detect and potentially deletes respondents whose scores are insufficiently reliable. The method builds on the strategy of presenting items within item triads.

As Thorndike and Hagen (1961, p. 27) noted long ago, "good measurement techniques provide the solid foundation for sound evaluation, whether of a single pupil or of a total curriculum." Fortunately, recent advances have encouraged evaluators to give more thought to measurement aspects of their evaluation designs. For example, new developments in latent trait theory have stimulated a substantial rethinking of measurement issues (Hambleton, Swamiwathan, Cook, Eignor, and Gifford, 1978). An unrelated development which has also provided an impetus to thought in this area has been the increasing recognition that analysis of covariance will not always magically equalize non-equivalent treatment groups (Campbell and Erlebacher, 1975). This realization has led to more widespread use of normal curve equivalent scores and the development of a set of mandated models for evaluating Title I programs (see EEPA, March-April, 1979).

Despite these developments, there is still considerable room for improvement in measurement practice in evaluation. For example, "response error" may pose a serious unrecognized threat to the external validity of some evaluation studies. Nunnally (1967, pp. 594-595) has suggested that "although the names frequently are used interchangeably in the literature, it is important to make a careful distinction between 'response styles' and 'response bias.'" Response styles refer to reliable response patterns of individuals as individuals; response bias refers to response patterns of groups which distort averages. Taken together, these two dynamics may be considered "response error" influences.

Several kinds of response styles can be identified, but probably the most common source of response style variance involves the yea-saying or nay-saying tendancies of some individuals (Schultz and Foster, 1963). Similarly, several sorts of response biases can be identified, but probably the most common source involves sensitivity to the social desirability of responses (Edwards, 1961). Response error distortions of both types are particularly likely to be evident in measurements for which there are no "one right" answers (Cronbach, 1950). Of course, there are no "right answers" in many evaluation studies, with the noteable exception of studies which involve achievement tests. Achievement tests have historically been of great value in making evaluative decisions about educational programs, but other types of measures can be important, particularly if any of three situations apply.

First, measurement in the absence of "right answers" is unavoidable when the *affective* impacts of programs are examined (cf. Welch and Walberg, 1974). Affective measures simply do not have "right answers." It is important to assess the affective impacts of programs, because "schools do not (or at least should not) act within an intellectual vacuum unrelated to the affective goals that educators claim to be important" (Sax, 1974, p. 286).

Second, measurement in the absence of "right answers" is desirable when direct assessment of program impacts is not *practical* , and so evaluators ask program participants to self-report perceived program impacts or their attitudes toward a program. For example, the problems in experimentally identifying the impacts of inservice activities would generally be overwhelming, but if the desired impacts were particularly important or the activities were particularly expensive, the teachers might at least be asked to indicate their judgments of the activities.

Finally, measurement in the absence of "right answers" is unavoidable when it is necessary to maximize the *credibility* of an evaluation study. Stufflebeam, Foley, Gephart, Guba, Hammond, Merriman, and Provus (1971) have argued that evaluation research should be both technically sound and credible to evaluation clients. Studies which are not credible to clients can hardly be expected to influence practice, notwithstanding their technical merits. Involving program participants in the evaluation of programs can provide an effective vehicle for maximizing evaluation credibility. When the number of participants is large, this is best accomplished with the use of questionnaires or surveys.

In short, the preceeding analysis suggests that response error can pose a threat to the external validity of some evaluation studies. The purpose of this paper is to assess how serious the threat may be, and to discuss one possible strategy for minimizing this source of error variance.

Magnitude of the Threat

In order to accomplish the paper's first purpose, three different sets of evaluation data were examined. Each set of data involved a "no right answers" situation. The data were collected in a large urban school district located in the southwestern United States. The first data set consisted of responses by 2466 teachers to a questionnaire which focused on perceptions of a drug-abuse inservice program. The questionnaire consisted of six items; each item had five Likert-type response alternatives. The second data set consisted of responses by 1657 teachers to a questionnaire which focused on perceptions of a special education inservice program. The questionnaire consisted of seven items; each item had five Likert-type response alternatives. The third data set consisted of responses by 243 teachers to a questionnaire which focused on perceptions of an inservice program for teachers of disadvantaged students. The questionnaire consisted of 15 items; each item had four Likert-type response alternatives.

The vast preponderance of responses to all items across all three instruments were at the "socially desirable" ends of the Likert-scales. On the average, less than 10% of the respondents selected responses toward the less socially desirable end of the scales. This result may or may not have reflected the influence of response bias. Perhaps all three programs were of uniformly high quality across various quality criteria.

The second step in the analysis was performed in order to estimate the seriousness of response style impacts. The data were analyzed to identify patterns in the responses of individual respondents. For the first data set, 560 teachers (22.7%) answered at least five of the six items identically. That is, if one of these teachers selected response alternative two for item one, response alternative two was selected for at least four of the remaining five items. For the second data set, 433 teachers (26.2%) answered at least five of the seven questions identically.

For the third data set, 101 teachers (41.6%) answered all 15 of the 15 questions identically!

These results suggest that response style can seriously confound the interpretation of some evaluation results. The results may reflect a disposition of some respondents to answer questionnaire items in terms of a global impression of the inservice activities. At any rate, these analyses suggest that the minimization and detection of response error ought to be of concern in certain evaluation designs.

A Possible Solution

Efforts to identify strategies for minimizing and detecting response error must be based on some theory about how the phenomenon functions. Scott (1968-a, p. 236) has suggested that both subject and instrument factors contribute to response error, although "the distinction between subject factors and instrument factors is a hard one to draw..., and some would assert that all response tendancies depend on the interaction between subject and instrumentation."

Probably the most influential subject factors which contribute to response error in evaluation data are situational. In some cases responses are influenced by respondent perceptions that "they don't look at these things anyway" or that "my one voice among many can't be heard anyway." This suggests a somewhat tragic paradox. Response error is most likely to be problematic when "no one right answer" evaluation data are especially needed, i.e.-- when comprehensive discussions between program clients and decision-makers are impractical due to number or time constraints. Fortunately, although these subject factors are difficult to modify, "much can be done by appropriate design of the survey and analysis of the responses themselves" (O'Muircheartaigh, 1977, p. 206).

One strategy for minimizing and detecting response error involves the use of a forced-choice item format. Table 1 presents data provided by one subject in a descriptive evaluation of the value orientation of a curriculum project. The subjects were asked to rank the items in each triad according to how strongly the project curriculum emphasized each value. Of course, several other forced-choice formats can be identified, but the strategy presented here is useful because it allows estimation of *intra*-subject reliabilities for each respondent.

Forced-choice formats were originally developed to minimize response error in the evaluation of Army personnel. Likert-scale evaluations tended to be very negatively skewed, so discrimination as to which officers were particularly capable was difficult to obtain (Sisson, 1948).

Zavala (1965, p. 117) has suggested that "studies on the FC [forced-choice] method show that this scale is more resistant than other scales to effects of bias." Scott (1968-b) argued that the format yields reliability and validity coefficients which are at least as good as those generated by other techniques, even when response error tendancies are not particularly strong.

The forced-choice format has been criticized by some for being artificial. However, Kerlinger (1973, p. 507) has argued that such choosing "is really a customary human activity... It can even be argued that agreement-disagreement items are artificial and that choice items are 'natural.'" In any case, the format is perfectly direct, i.e.--non-artificial, when the evaluation asks, "Given that most of these statements may be basically true, which ones are most true?" Discussion of the technique will be couched in terms of a five item instrument. However, the logic generalizes to other situations.

<u>Insert Table 1 about here.</u>

The data presented in Table 1 can be re-arranged into the format presented in Table 2. Table 2 indicates that a five item forced-choice triad measurement involves 10 pair-wise comparison judgments, e.g.-- A>B, A>C, etc. Table 2 also indicates that each pair-wise judgment for a five

TABLE I

One Subject's Triad Responses

Response	Item	Item Name	Order	Response	Item	Item Name	Order
1	Honesty	(A)	(9-2)	1	Honesty	(A)	(10-1)
2	Money	(B)	(9-1)	3	Fame	(D)	(10-3)
3	Love	(C)	(9-3)	2	Power	(E)	(10-2)
1	Honesty	(A)	(1-1)	1	Money	(B)	(2-2)
3	Money	(B)	(1-2)	3	Love	(C)	(2-1)
2	Fame	(D)	(1-3)	2	Fame	(D)	(2-3)
1	Honesty	(A)	(4-1)	1	Money	(B)	(8-3)
2	Money	(B)	(4-2)	2	Love	(C)	(8-2)
3	Power	(E)	(4-3)	3	Power	(E)	(8-1)
1	Honesty	(A)	(6-3)	1	Money	(B)	(7-2)
2	Love	(C)	(6-2)	2	Fame	(D)	(7-1)
3	Fame	(D)	(6-1)	3	Power	(E)	(7-3)
1	Honesty	(A)	(5-2)	3	Love	(C)	(3-3)
2	Love	(C)	(5-1)	1	Fame	(D)	(3-1)
3	Power	(E)	(5-3)	2	Power	(E)	(3-2)

NOTE: Each of the items was assigned a name, A, B, C, D, or E, in order to facilitate discussion in the narrative. "Order" indicates respectively the original order of the triad among the 10 triads and then item order within each triad.

item forced-choice triad measurement is replicated three times, e.g.-- ABC, ABD, and ABE. The notion of replication suggests that a test-retest reliability coefficient could be generated for this individual. The calculated intra-subject reliability for these data was .91. This value was attenuated in this case by the inconsistent judgments made for item pairs ABD, ACD, CDE, and ADE.

Insert Table 2 about here.

Subjects whose responses are seriously inconsistent, i.e.-- have especially low intra-subject reliabilitity estimates, might be deleted from the analysis. A pooled reliability estimate for a data set can be calculated by averaging these values. Finally, the strategy tends to maximize the reliability of results since there are several repitions of each judgment (Remmers, Shock, and Kelly, 1927). Once the data have been examined for reliability, they can be aggregated into rankings or other values, and then medians or other solutions can be computed (cf. Coxon and Jones, 1977).

Of course, the calculations involved in this approach could be quite staggering. This would especially be true if item triads were randomly ordered and items were also randomly ordered within triads. Such random ordering itself tends to minimize response errors. However, a computer

TABLE II

Matrix for Estimating Intra-Subject Reliability

				Replications I to III					
Judgment	I	Source	Data	II	Source	Data	III	Source	Data
A>B?		ABC	Yes=1		ABD	Yes=1		ABE	Yes=1
A>C?		ACD	Yes=1		ACE	Yes=1		ABC	Yes=1
A>D?		ADE	Yes=1		ACD	Yes=1		ABD	Yes=1
A>E?		ABE	Yes=1		ADE	Yes=1		ACE	Yes=1
B>C?		BCD	Yes=1		ABC	Yes=1		BCE	Yes=1
B>D?		ABD	No=0*		BDE	Yes=1		BCD	Yes=1
B>E?		BCE	Yes=1		ABE	Yes=1		BDE	Yes=1
C>D?		CDE	No=0		BCD	No=0		ACD	Yes=1*
C>E?		ACE	Yes=1		BCE	Yes=1		CDE	No=0*
D>E?		BDE	Yes=1		CDE	Yes=1		ADE	No=0*

Note: The judgments which appear to be internally inconsistent have been highlighted with an asterisk.

program which performs the necessary calculations for randomly ordered 10 triad (five item) and 20 triad (6 item) instruments is available at no cost from the senior author.

Conclusions

These analyses have at least three implications for evaluation practice. The results make clear that some evaluation data are highly skewed. This is serious because decision-makers tend to interpret measures of central tendancy by comparing them with scale mid-point values. This presumes that the latent "true distributions" underlying the measures are symetrical so that the expected mean or median in a "no one right answer" situation actually would be the scale mid-point. Evaluators have some obligation to emphasize to decision-makers that such comparisons may not be legitimate. Even responses which are more favorable than a scale mid-point may in some cases still be relatively or even dramatically negative.

The results also indicate that evaluators ought to check for some forms of response style in some "no one right answer" situations, when Likert-scale measurements are made. The pattern of an individual answering every item identically can be readily identified with statistics packages which employ IF and COUNT cards.

Finally, the analysis indicates that forced-choice strategies may be helpful in some "no one right answer" evaluation situations. The strategies may be particularly helpful when respondents feel distant from desicion-makers. Also, the strategies may be helpful when evaluators focus on attitude changes as program impacts, because response errors tend be compounded when other formats are used and both pre and post measures are taken (Bartlett, Quay, and Wrightsman, 1960).

Several strategies for reducing response error are available to evaluators. However, the strategy proposed herein may be particularly valuable to the extent that it provides estimates of intra-subject reliability. The importance of these estimates will be a function of how critical it is to detect response error in any given evaluation study. In any case, as Lansing, Ginsburg, and Braaten (1961, p. 7) have noted, "the problem of response error is quite serious, and in part one must learn to live with it; but there are ways of controlling it and researchers certainly can work to minimize it."

References

Bartlett, C.J., Quay, L.C., & Wrightsman, L.S.J. A comparison of two methods of attitude measurement: Likert-type and forced choice. Educational and Psychological Measurement, 1960, 20, 699-704.

Campbell, D.T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In E.L. Struening & M. Guttentag (Eds.), Handbook of evaluation research (Vol. 1). Beverly Hills: Sage, 1975.

Coxon, A.P.M., & Jones, C.L. Multidimensional scaling. In C.A. O'Muircheartaigh & C. Payne (Eds.), Exploring data structures (Vol. 1). New York: John Wiley and Sons, 1977.

Cronbach, L.J. Further evidence on response sets and test design. Educational and Psychological Measurement, 1950, 10, 3-31.

Edwards, A.L. Social desirability or acquiescence in the MMPI: A case study with the SD scale. Journal of Abnormal and Social Psychology, 1961, 63, 351-359.

Hambleton, R.K., Swamiwathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Kerlinger, F.N. Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart, and Winston, 1973.

Lansing, J.B., Ginsburg, G.P., & Braaten, K. An investigation of response error. Urbana: Bureau of Economic and Business Research, 1961.

Nunnally, J.C. Psychometric theory. New York: McGraw-Hill, 1967.

O'Muircheartaigh, C.A. Response errors. In C.A. O'Muircheartaigh & C. Payne (Eds.), The analysis of survey data (Vol. 2). New York: John Wiley, 1977.

Remmers, H.H., Shock, N.W., & Kelly, E.L. An empirical study of the validity of the Spearman Brown formula as applied to the Purdue Rating Scale. Journal of Educational Psychology, 1927, 18, 187-195.

Sax, G. The use of standardized tests in evaluation. In W.J. Popham (Ed.), Evaluation in education. Berkeley: McCutchan, 1974.

Schutz, R.G., Foster, R.J. A factor analytic study of acquiescent and extreme response set. Educational and Psychological Measurement, 1963, 23, 435-447.

Scott, W. Attitude measurement. In G. Lindzey and E. Aronson (Eds.), *The handbook of social psychology* (2nd ed., Vol. 2). Reading, MS: Addison-Wesley, 1968-a.

Scott, W. Comparative validities of forced-choice and single-stimulus tests. *Psychological Bulletin*, 1968-b, *70*, 231-244.

Sisson, E.D. Forced choice--The new Army rating. *Personnel Psychology*, 1948, *1*, 365-381.

Stufflebeam, D., Foley, W.J., Gephart, W.J., Guba, E.G., Hammond, R.L., Merriman, H.D., & Provus, N.M. *Educational evaluation and decision making*. Itasca, IL: Peacock, 1971.

Thorndike, R.L., & Hagen, E. *Measurement and evaluation in psychology and education* (2nd ed.). New York: John Wiley and Sons, 1961.

Welch, W.W., & Walberg, H.J. A course evaluation. In H.J. Walberg (Ed.), *Evaluating educational performance*. Berkeley: McCutchan, 1974.

Zavala, A. Development of the forced-choice rating scale technique. *Psychological Bulletin*, 1965, *63*, 117-124.