2003 Midwest Research to Practice Conference in Adult, Continuing, and Community Education

Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales

Joseph A. Gliem
Rosemary R. Gliem

Abstract: The purpose of this paper is to show why single-item questions pertaining to a construct are not reliable and should not be used in drawing conclusions. By comparing the reliability of a summated, multi-item scale versus a single-item question, the authors show how unreliable a single item is; and therefore it is not appropriate to make inferences based upon the analysis of single-item questions which are used in measuring a construct.

Introduction

Oftentimes information gathered in the social sciences, marketing, medicine, and business, relative to attitudes, emotions, opinions, personalities, and description's of people's environment involves the use of Likert-type scales. As individuals attempt to quantify constructs which are not directly measurable they oftentimes use multiple-item scales and summated ratings to quantify the construct(s) of interest. The Likert scale's invention is attributed to Rensis Likert (1931), who described this technique for the assessment of attitudes.

McIver and Carmines (1981) describe the Likert scale as follows:

> A set of items, composed of approximately an equal number of favorable and unfavorable statements concerning the attitude object, is given to a group of subjects. They are asked to respond to each statement in terms of their own degree of agreement or disagreement. Typically, they are instructed to select one of five responses: strongly agree, agree, undecided, disagree, or strongly disagree. The specific responses to the items are combined so that individuals with the most favorable attitudes will have the highest scores while individuals with the least favorable (or unfavorable) attitudes will have the lowest scores. While not all summated scales are created according to Likert's specific procedures, all such scales share the basic logic associated with Likert scaling. (pp. 22-23)

Spector (1992) identified four characteristics that make a scale a summated rating scale as follows:

> First, a scale must contain multiple items. The use of *summated* in the name implies that multiple items will be combined or summed. Second, each individual item must measure something that has an underlying, quantitative measurement continuum. In other words, it measures a property of something that can vary quantitatively rather than qualitatively.

> An attitude, for example, can vary from being very favorable to being very unfavorable. Third, each item has no "right" answer, which makes the summated rating scale different from a multiple-choice test. Thus summated rating scales cannot be used to test for knowledge or ability. Finally, each item in a scale is a statement, and respondents are asked to give rating about each statement. This involves asking subjects to indicate which of several response choices best reflects their response to the item. (pp. 1-2)

Nunnally and Bernstein (1994), McIver and Carmines (1981), and Spector (1992) discuss the reasons for using multi-item measures instead of a single item for measuring psychological attributes. They identify the following: First, individual items have considerable random measurement error, i.e. are unreliable. Nunnally and Bernstein (1994) state, "Measurement error averages out when individual scores are summed to obtain a total score" (p. 67). Second, an individual item can only categorize people into a relatively small number of groups. An individual item cannot discriminate among fine degrees of an attribute. For example, with a dichotomously scored item one can only distinguish between two levels of the attribute, i.e. they lack precision. Third, individual items lack scope. McIver and Carmines (1981) say, "It is very unlikely that a single item can fully represent a complex theoretical concept or any specific attribute for that matter" (p. 15). They go on to say,

> The most fundamental problem with single item measures is not merely that they tend to be less valid, less accurate, and less reliable than their multiitem equivalents. It is rather, that the social scientist rarely has sufficient information to estimate their measurement properties. Thus their degree of validity, accuracy, and reliability is often unknowable. (p. 15).

Blalock (1970) has observed, "With a single measure of each variable, one can remain blissfully unaware of the possibility of measurement [error], but in no sense will this make his inferences more valid" (p. 111).

Given this brief background on the benefits of Likert-type scales with their associated multi-item scales and summated rating scores, many individuals consistently invalidate research findings due to improper data analysis. This paper will show how data analysis errors can adversely affect the inferences one wishes to make.

Data Analysis Errors with Likert Scales

Reporting Errors with Reliability Measures

While most individuals utilizing Likert-type scales will report overall scale and subscale internal consistency reliability estimates in the analysis of the data, many will analyze individual scale items. Table 1 shows the results of an analysis done be Warmbrod (2001) of *The Journal of Agricultural Education*, 2000 (Volume 41). Volume 41 of the journal contained 44 articles of which 36 (82%) were quantitative. Of these 36 articles, 29 (66%) used researcher developed Likert scales for which internal consistency reliability measures need to be reported. The table shows the reliability coefficients reported and the analysis strategy used with the quantitative articles contained in the journal. As shown in the table, only two (7%) of the individuals correctly analyzed the data collected based upon the reliability measures reported. The majority of individuals correctly reported Cronbach's alpha as the measure of internal consistency

reliability, but then opted to conduct data analysis using individual items. This is particularly troubling because single item reliabilities are generally very low, and without reliable items the validity of the item is poor at best and at worst unknown. This can be illustrated using a simple data set of actual data collected from a class of graduate students enrolled in a Winter Quarter, 2003, research design course. Cronbach's alpha is a test reliability technique that requires only a single test administration to provide a unique estimate of the reliability for a given test. Cronbach's alpha is the average value of the reliability coefficients one would obtained for all possible combinations of items when split into two half-tests.

Table 1: Reliability estimates and analysis strategies: Researcher-developed multiple-item instruments with Likert-type scaling

Number	%	Reliability Coefficients Reported	Analysis Strategy
3	10.3	None	Single item analysis exclusively
14	48.3	Chronbach's alpha: Total scale and/or subscales	
1	3.4	Chronbach's alpha: Total scale and/or subscales Test-Retest: selected items	Single item analysis exclusively
1	3.4	Chronbach's alpha: Total scale and/or subscales	Single item analysis exclusively Some summated score analysis
4	13.8	Chronbach's alpha: Total scale and/or subscales	Single item analysis and summated score analysis
4	13.8	Chronbach's alpha: Total scale and/or subscales	Summated score analysis primarily; some single item analysis
2	7.0	Chronbach's alpha: Total scale and/or subscales	Summated score analysis exclusively

Calculating Cronbach's Alpha Coefficient for Internal Consistency

A single statement (item) was presented to each student and then this same statement was presented to the student 3 weeks later. A test-retest reliability coefficient was calculated on this individual statement (item) since individual items can not have a Cronbach's alpha internal consistency reliability calculated. The statement presented to each student was, "I am pleased with my graduate program at The Ohio State University." Students were asked to respond to the statement using a five-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). Figure 1 shows a scatterplot of student response for the first administration of the statement and for the second administration of the statement 3 weeks later. The test-retest reliability coefficient for this statement was .11. A multi-item scale was also developed and given to the same students to measure their attitude towards their graduate program. The multi-item scale is presented in Table 2.

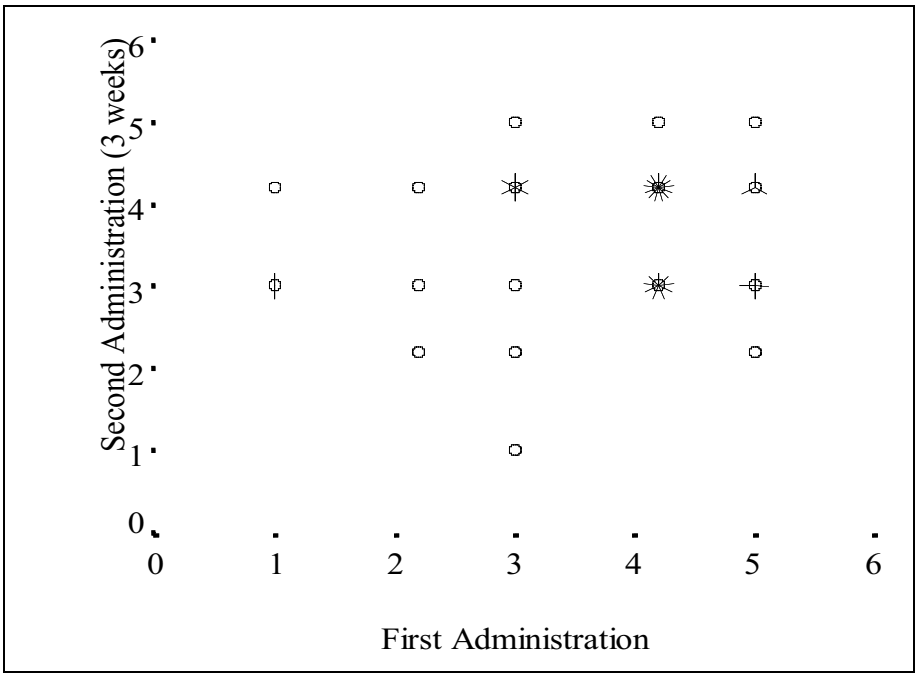

Figure 1: Scatterplot of first administration and second administration; r = .11

Table 2: Multi-item statements to measure students pleasure with their graduate program at The Ohio State University

Item	Strongly Disagree				Strongly Agree
1. My advisor is knowledgeable about my program	1	2	3	4	5
2. It is easy to get an appointment with my advisor	1	2	3	4	5
3. My advisor cares about me as a person	1	2	3	4	5
4. My advisor considers me a professional	1	2	3	4	5
5. The course requirement for my major are appropriate	1	2	3	4	5
6. The course requirements for my major will help me get a job in my discipline	1	2	3	4	5
7. My graduate program allows me adequate flexibility	1	2	3	4	5
8. Upon graduation, I anticipate I will not have any problems finding a position in my discipline	1	2	3	4	5
9. My graduate program needs to be updated	1	2	3	4	5
10. The quality of the courses required in my major is adequate	1	2	3	4	5

Table 3 shows the item-analysis output from SPSS for the multi-item scale of student attitude towards their graduate program. A description of the sections and related terms are as follows:

1. Statistics for Scale—These are summary statistics for the 8 items comprising the scale. The summated scores can range from a low of 8 to a high of 40.
2. Item means—These are summary statistics for the eight individual item means.
3. Item Variances—These are summary statistics for the eight individual item variances.
4. Inter-Item Correlations—This is descriptive information about the correlation of each item with the sum of all remaining items. In the example in Table 2, there are 8 correlations computed: the correlation between the first item and the sum of the other seven items, the correlation between the second item and the sum of the other seven items, and so forth. The first number listed is the mean of these eight correlations (in our example .3824), the second number is the lowest of the eight (.0415), and so forth. The mean of the inter-item correlations (.3824) is the r in the _ = rk / [1 + (k -1) r] formula where k is the number of items considered.
5. Item-total Statistics—This is the section where one needs to direct primary attention. The items in this section are as follows:
    a. Scale Mean if Item Deleted—Excluding the individual item listed, all other scale items are summed for all individuals (48 in our example) and the mean of the summated items is given. In Table 2, the mean of the summated scores excluding item 2 is 25.1.
    b. Scale Variance if Item Deleted—Excluding the individual item listed, all other scale items are summed for all individuals (48 in our example) and the variance of the summated items is given. In Table 2, the variance of the summated scores excluding item 2 is 25.04.
    c. Corrected Item-Total Correlation—This is the correlation of the item designated with the summated score for all other items. In Table 2, the correlation between item 2 and the summated score is .60. A rule-of-thumb is that these values should be at least .40.
    d. Squared Multiple Correlation—This is the predicted Multiple Correlation Coefficient squared obtained by regressing the identified individual item on all the remaining items. In Table 2, the predicted Squared Multiple Regression Correlation is .49 by regressing item 2 on items 4, 5, 6, 7, 8, 9, and 10.
    e. Alpha if Item Deleted—This is probably the most important column in the table. This represents the scale's Cronbach's alpha reliability coefficient for internal consistency if the individual item is removed from the scale. In Table 2, the scale's Cronbach's alpha would be .7988 if item 2 were removed for the scale. This value is then compared to the Alpha coefficient value at the bottom of the table to see if one wants to delete the item. As one might have noted, the present scale has only 8 items where the original scale had 10 items. Using the above information, removing items 1 and 2 resulted in an increase in Cronbach's alpha from .7708 to .8240.
    f. Alpha—The Cronbach's alpha coefficient of internal consistency. This is the most frequently used Cronbach's alpha coefficient.
    g. Standardized Item Alpha—The Cronbach's alpha coefficient of internal consistency when all scale items have been standardized. This coefficient is used only when the individual scale items are not scaled the same.

Table 3: Item-Analysis From SPSS Output

	N	Mean	Variance	SD
Statistics for Scale	8	29.1042	30.8187	5.5515

	Mean	Minimum	Maximum	Range	Max/Min	Variance
Item Means	3.6380	3.3125	3.9792	.6667	1.2013	.0729
Item Variances	1.0750	.7017	1.4109	.7092	2.0107	.0714
Inter-Item Correlations	.3824	.0415	.5861	.5446	14.1266	.0191

Item Total Statistics	Scale Mean If Item Deleted	Scale Variance If Item Deleted	Corrected Item Total Correlation	Squared Multiple Correlation	Alpha If Item Deleted
Item 2	25.1250	25.0479	.6046	.4909	.7988
Item 4	25.7917	23.2748	.5351	.3693	.8063
Item 5	25.6667	24.6525	.4260	.4474	.8219
Item 6	25.2500	25.2128	.5134	.4587	.8081
Item 7	25.6250	22.9202	.6578	.5104	.7874
Item 8	25.7083	24.3387	.4473	.3116	.8192
Item 9	25.1250	23.9840	.6134	.5202	.7949
Item 10	25.4375	24.0811	.6432	.4751	.7920

	Alpha	Standardized Item Alpha
Reliability Coefficients for Item 8	.8240	.8320

Cronbach's alpha reliability coefficient normally ranges between 0 and 1. However, there is actually no lower limit to the coefficient. The closer Cronbach's alpha coefficient is to 1.0 the greater the internal consistency of the items in the scale. Based upon the formula _ = rk / [1 + (k -1)r] where k is the number of items considered and r is the mean of the inter-item correlations the size of alpha is determined by both the number of items in the scale and the mean inter-item correlations. George and Mallery (2003) provide the following rules of thumb: "_ > .9 – Excellent, _ > .8 – Good, _ > .7 – Acceptable, _ > .6 – Questionable, _ > .5 – Poor, and _ < .5 – Unacceptable" (p. 231). While increasing the value of alpha is partially dependent upon the number of items in the scale, it should be noted that this has diminishing returns. It should also be noted that an alpha of .8 is probably a reasonable goal. It should also be noted that while a high value for Cronbach's alpha indicates good internal consistency of the items in the scale, it does not mean that the scale is unidimensional. Factor analysis is a method to determine the dimensionality of a scale but is beyond the scope of this paper.

Conclusions

When using Likert-type scales it is imperative to calculate and report Cronbach's alpha coefficient for internal consistency reliability for any scales or subscales one may be using. The analysis of the data then must use these summated scales or subscales and not individual items. If one does otherwise, the reliability of the items is at best probably low and at worst unknown. Cronbach's alpha does not provide reliability estimates for single items.

References

Blalock, H. M., Jr. (1970). Estimating measurement error using multiple indicators and several points in time. *American Sociological Review*, *35*(1), 101-111.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks, CA: Sage.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

Likert, R. (1931). A technique for the measurement of attitudes. *Archives of Psychology*. New York: Columbia University Press.

McIver, J. P., & Carmines, E. G. (1981). *Unidimensional scaling*. Thousand Oaks, CA: Sage.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Spector, P. (1992). *Summated rating scale construction*. Thousand Oaks, CA: Sage.

Warmbrod, J. R. (2001). *Conducting, interpreting, and reporting quantitative research*. Research Pre-Session, New Orleans, Louisiana.

Joseph A. Gliem, Associate Professor, Dept. of Human and Community Resource Development, The Ohio State University, 208 Ag. Admin Bldg., 2120 Fyffe Rd., Columbus, OH 43210; gliem.2@osu.edu

Rosemary R. Gliem, The Ohio State University; gliem.1@osu.edu

Presented at the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, The Ohio State University, Columbus, OH, October 8-10, 2003.