Journal of Agricultural Education, 55(5), 30-47.
doi: 10.5032/jae.2014.05030

Reporting and Interpreting Scores Derived from Likert-type Scales

J. Robert Warmbrod[1]

Abstract

Forty-nine percent of the 706 articles published in the Journal of Agricultural Education *from 1995 to 2012 reported quantitative research with at least one variable measured by a Likert-type scale. Grounded in the classical test theory definition of reliability and the tenets basic to Likert-scale measurement methodology, for the target population of 344 articles using Likert-scale methodology, the objectives of the research were to (a) describe the scores derived from Likert-type scales reported and interpreted, (b) describe the reliability coefficients cited for the scores interpreted, and (c) ascertain whether there is congruence or incongruence between the reliability coefficient cited and the Likert-scale scores reported and interpreted. Twenty-eight percent of the 344 articles exhibited congruent interpretations of Likert-scale scores, 45% of the articles exhibited incongruent interpretations, and 27% of the articles exhibited both congruent and incongruent interpretations. Single-item scores were reported and interpreted in 63% of the articles, 98% of which were incongruent interpretations. Summated scores were reported and interpreted in 59% of the articles, 91% of which were congruent interpretations. Recommendations for analysis, interpretation, and reporting of scores derived from Likert-type scales are presented.*

Keywords: Reliability; Likert-type scale; Cronbach's alpha

During the 18-year period 1995 to 2012, 706 articles were published in the *Journal of Agricultural Education*. Forty-nine percent of the 706 articles were reports of quantitative research with at least one variable measured by a Likert-type scale. Likert-scale methodology was used in 62% of the articles reporting quantitative research (see Table 1).

Grounded by the rationale and principles basic to the quantification of constructs using Likert-type scales and the theory of reliability of measurement, this article reports an investigation of the extent scores derived from Likert-type scales reported in the *Journal of Agricultural Education* are congruent with the estimates of reliability of measurement cited in the articles. The article deals exclusively with the reliability of test scores derived from a Likert-type scale. Equally important, but not addressed in the article, is evidence researchers present in journal articles documenting the validity of test scores, including a description of item-generating strategies to establish content validity, judgments of experts attesting face validity, and empirical evidence documenting criterion and construct validity (Nunnally & Bernstein, 1994, Chapter 3).

Principles underlying the research reported in the article are (a) reliability of measurement is a property of the *test scores* derived from the measurement instrument and (b) the standards for reporting research require authors to cite reliability coefficients for the test scores that are reported and interpreted (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Wilkinson & The Task Force on Statistical Inference, 1999). When authors fail to cite reliability coefficients for test scores or cite reliability coefficients incongruent with the test scores reported

[1] J. Robert Warmbrod is Distinguished University Professor Emeritus in the Department of Agricultural Communication, Education, and Leadership at the Ohio State University, 208 Agricultural Administration Building, 2120 Fyffe Road, Columbus, OH. Email: warmbrod.1@osu.edu.

and interpreted, evidence documenting the accuracy of measurement for the variables being investigated is unknown, thereby violating a basic standard for reporting educational and psychological test results.

Table 1

Articles Published in the Journal of Agricultural Education: 1995 – 2012

Articles published: 1995 – 2012	No. of articles	% of 706 articles	% of 554 articles
Total articles published	706	100.0	
Articles reporting non-quantitative research[a]	152	21.5	---
Articles reporting quantitative research	554	78.5	100.0
Articles with no Likert-type scale	210	29.8	37.9
→ Articles with Likert-type scale	344	48.7	62.1

[a] AAAE Distinguished Lecture, review and synthesis of research, historical research, philosophical research, content analysis, and qualitative research.

The Likert Scale

More than 80 years ago psychologist Rensis Likert published a monograph, *A Technique for the Measurement of Attitudes*, describing the concepts, principles, and substantiative research basic to an instrument to quantify constructs describing psychological and social phenomena (Likert, 1932). A Likert-type scale consists of a series of statements that define and describe the content and meaning of the construct measured. The statements comprising the scale express a belief, preference, judgment, or opinion. The statements are composed to define collectively an unidimensional construct (Babbie, 1999; McIver & Carmines, 1981). Alternatively, clusters of statements within a scale may define one or more subscales that quantify more specific unidemensional subconstructs within the major scale. In designing a Likert scale, the generation and wording of individual statements are crucial tasks for producing an instrument that yields valid and reliable summated scores (Edwards, 1957; Oppenheim, 1992; Spector, 1992).

The response continuum for each statement is a linear scale indicating the extent respondents agree or disagree with each statement. For example, a generic response continuum is 1 = Strongly Disagree, 2 = Disagree, 3 = Undecided or Neutral, 4 = Agree, and 5 = Strongly Agree for statements favorable to the construct. For statements unfavorable to the construct – negatively worded statements – the numerical values for the response options are reversed when the summated score for the construct is calculated.

Likert's (1932) monograph specifies that the quantification of the construct is a summated score for each individual calculated by summing an individual's responses for each item comprising the scale. Kerlinger (1986) described a Likert scale as a summated rating scale whereby an inividual's score on the scale is a sum, or average, of the individual's responses to the multiple items on the instrument. Oppenheim (1992), Kline (1998), and Babbie (1999) emphasized that the score an individual receives on a Likert scale is the sum of an individual's responses to all items comprising the scale or subscale. *A principle basic to Likert scale measurement methodology is that scores yielded by a Likert scale are composite (summated) scores derived from an individual's responses to the multiple items on the scale.*

An alternative procedure for calculating a composite score for each individual is to calculate a mean-item summated score, that is, an individual's summated score divided by the number of items constituting the scale or subscale thereby creating a mean-item score for each individual that falls within the range of the values for the response continuum options. All items comprising a scale or subscale are assumed to have equal weight when calculating a summated score or a mean-item score.

The content of single items (statements) on a Likert scale collectively define, describe, and name the meaning of the construct quantified by the summated score. When reporting research it is appropriate to list the statements that define the unidemensional construct and record the percentage of respondents choosing each response option. These summary statistics for each item on the scale indicate the content of the construct and the direction and intensity of each item's contribution to the summated total score or summated subscale score.

Two basic concepts provide the rationale for reporting and interpreting summated scores derived from Likert-type scales to quantify psychological, sociological, and educational constructs. First is the proposition that the construct being measured is not defined by a single statement. A Likert scale is by definition a multiple-item scale. The second defining characteristic logically follows: *scores derived from a Likert scale are summated scores determined by a composite of responses to multiple items rather than responses to single items.*

McIver and Carmines (1981), Nunnally and Bernstein (1994), and Oppenheim (1992) contended it is unlikely that a single item can adequately represent a complex underlying construct. Hair, Anderson, Tatham, and Black (1998) emphasized that using responses to a single item as representative of a concept runs the risk of potentially misleading results by selecting a single statement to represent a more complex result. Responses to single items usually have a low degree of relationship with a composite score derived from responses to multiple items defining the construct.

Measurement specialists (McIver & Carmines, 1981; Nunnally & Bernstein, 1994) reported that single items tend to be less valid, less accurate, and less reliable than multiple-item composites; that responses to single items have considerable measurement error; and that sufficient information is rarely available to estimate the accuracy, validity, and reliability of a single item. The principle of aggregation – the sum of the responses to a set of multiple items is a more stable and unbiased estimate than are responses to any single item in the set – empirically demonstrates that summated scores derived from responses to multiple items on a Likert-type scale are more reliable than responses to single items comprising the scale (Rushton, Brainerd, & Pressley, 1983; Strube, 2000). Classical test theory assumes random error is always associated with measurement. When responses to the set of single items defining a construct are combined, the random measurement errors tend to average out thereby providing a more reliable composite measure of the construct. Blalock's (1970) investigation of the single-item versus multiple-item issue concluded with these statements: "With a single measure of each variable, one can remain blissfully unaware of the possibility of measurement error. I see no substitute for the use of multiple measures of our most important variables" (p. 111).

Researchers in agricultural education use Likert-type scales to measure attitudes about policies and programs regarding education in and about agriculture; perceptions of barriers, benefits, and challenges to practices and programs; teacher efficacy; job satisfaction; and self-perceptions of level of knowledge and competence. Table 2 lists examples of articles published in the *Journal of Agricultural Education* where Likert-type scales were used to quantify constructs.

Table 2

Examples of Constructs Measured in Articles Published in the Journal of Agricultural Education

Example 1	
Construct:	Teacher Efficacy – Overall efficacy (24 items); Student engagement subscale (8 items); Instructional strategies subscale (8 items); Classroom management subscale (8 items)
Response continuum:	How much can you do? 1 = Nothing, 3 = Very little, 5 = Some influence, 7 = Quite a bit, 9 = A great deal
Target population:	Agricultural science student teachers
Example 2	
Construct:	Attitude toward agriculture (13 items)
Response continuum:	0 = Strongly disagree, 1 = Disagree, 2 = Neutral, 3 = Agree, 4 = Strongly agree
Target population:	Secondary school students enrolled in agriscience courses
Example 3	
Construct:	Perception concerning the integration of instruction in science and agriculture (12 items)
Response continuum:	1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree
Target population:	Secondary school science teachers

The Concept of Reliability

Reliability describes the accuracy of measurement. Derived from classical test theory, the reliability of a test score that quantifies psychological and social constructs postutlates that an individual's true score is comprised of an observed (measured) score minus randon errors of measurement expressed by the the following equation (Cronbach, 1984, Chapter 6).

$$\text{True score} = \text{Observed score} - \text{Error} \qquad (1)$$

Applying this principle when a group of individuals has completed an instrument that measures a specific construct, it follows that the variance of the true scores for the group equals the variance of the group's observed scores minus the variance of the random errors of measurement (see Equation 2).

$$\text{Variance}_{(\text{True score})} = \text{Variance}_{(\text{Observed score})} - \text{Variance}_{(\text{Error})} \qquad (2)$$

When attitudinal and perceptual constructs are measured using Likert-type scales, an individual's observed score is a composite summated score, either a summated total score or a summated subscale score, which is the sum of an individual's responses to items comprising the Likert scale that define the construct being measured. In Equation 2, the variance of the observed summated scores is calculated for the group of individuals responding to the Likert scale. The variance of the errors of measurement, which are assumed to be random, are estimated from the variations among individuals by their responses to each item on the Likert scale. True scores on the construct being measured for individuals in the group are unknown; therefore, the variance of the true summated scores can only be estimated.

Reliability is expressed as a coefficient that is the proportion of the variance of the observed summated scores that is *not* attributed to random error variance, which is the ratio of estimated variance of unknown true scores to the calculated variance of the observed scores. This ratio is depicted in Equation 3.

$$\text{Reliability coefficient} = \frac{\text{Variance}_{(\text{True scores})}}{\text{Variance}_{(\text{Observed scores})}} \quad (3)$$

Because Equation 2 defines the variance of true scores as the variance of observed scores minus the variance of the random errors of measurement, the equation for estimating the reliability coefficient is presented in Equation 4.

$$\text{Reliability coefficient} = \frac{\text{Variance}_{(\text{Observed scores})} - \text{Variance}_{(\text{Errors})}}{\text{Variance}_{(\text{Observed scores})}} \quad (4)$$

The calculated reliability coefficient is an estimate because one term in the equation – variance of the random errors – is an estimate. When constructs are measured by Likert-type scales, statistically the reliability coefficient is an estimate of the proportion of variance in the observed summated scores that is *not* attributable to random errors of measurement. Values of the calculated reliability coefficient vary from 0.0 to 1.0 with values approaching 1.0 indicating that the observed summated scores are relatively free from random errors of measurement (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The reliability of a test score is frequently described as the dependability, consistency, or stability of the score produced by a particular instrument, which in this case is a summated total score or a summated subscale score derived from a Likert-type scale. An important point is that reliability of a score derived from a Likert scale is the *property of a summated score, not a characteristic of the instrument from which the summated score was derived*.

Estimating Reliability

Coefficient of stability. For a Likert-type scale, the test-retest procedure estimates reliability by calculating the correlation of summated scores administered to the same respondents on two different occasions. Estimating the reliability coefficient by the test-retest procedure requires consideration of two possible problems. First, if the time between the two administrations of the instrument is too short, the calculated correlation coefficient may be spuriously high due to the effect of recall, if on the second administration of the instrument, respondents remember how they responded on the first administration. Likewise, if the time between the two administrations of the instrument is too long, the calculated coefficient of stability may be low due to change on the part of respondents for the construct being measured. The longer the time between the two administrations, the more likely the construct has changed (Isaac and Michall, 1995).When the test-retest procedure is used to estimate reliability, the time between the two administrations of the Likert scale should be reported.

Internal consistency. An internal consistency estimate of the reliability of summated scores derived from a Likert scale requires only one administration of the instrument. Internal consistency refers to the extent to which there is cohesivness or inter-relatedness among the responses to the multiple items comprising the Likert scale. Cronbach (1951) developed this estimate of reliability and named the coefficient *alpha (α)*. The mathematical definition of

Cronbach's α is described by Equation 5 where n equals the number of items comprising the Likert scale.

$$\text{Cronbach's alpha} = \frac{n}{n-1}\left(1 - \frac{Variance_{(Error)}}{Variance_{(Observed\ scores)}}\right) \quad (5)$$

Equation 6 for computing Cronbach's α (Carmines & Zeller, 1979) indicates that the value of the internal consistency estimate of reliability of summated scores is determined by the mean of the inter-item correlations among responses to single items on the Likert scale and the number of items comprisig the scale. In Equation 6, *n* equals the number of items on the scale and *r* is the mean inter-item correlation.

$$\text{Cronbach's } \alpha = \frac{nr}{1 + r(n-1)} \quad (6)$$

The higher the mean inter-item correlation (*r*), the higher the value of Cronbach's α; the more items comprising the Likert scale, the higher the value of Cronbach's α.

Cronbach (1951) stated that the alpha coefficient is the mean of all possible split-half coefficients that can be calculated for a particular instrument. Carmines and Zeller (1979) reported that Cronbach's α is a conservative estimate of reliability.

Standards for Reliability

Table 3 reports standards proposed by measurement specialists for reporting and interpreting reliability coefficients.

Objectives of the Study

For the 344 articles reporting quantitative research published in the *Journal of Agricultural Education* from 1995 to 2012 with at least one variable measured by a Likert-type scale, the objectives of the study were:

1. Describe the scores derived from the Likert-type scales that are reported and interpreted.
2. Describe the reliability coefficients cited in the articles to accompany the Likert scale scores reported and interpreted.
3. Describe whether there is congruence or incongruence between the Likert-scale scores reported and interpreted and the reliability coefficients cited.

Procedure

For each of the 344 articles in the target population of articles, the author reviewed each article and recorded a Yes(1) or No(0) response for the following questions.

1. Are scores for responses to single items on the Likert-type scale reported and interpreted?
2. Is a summated total score for the multiple items comprising the Likert-type scale reported and interpreted?
3. Are summated subscale scores for clusters of items within the Likert-type scale reported and interpreted?
4. Are reliability coefficients cited for single-item scores? If yes, what coefficient is cited?
5. Is a reliability coefficient cited for a summated total score? If yes, what coefficient is cited?
6. Are reliability coefficients cited for summated subscale scores? If yes, what coefficient is cited?

Responses to the six questions were used to categorize each article on two factors: *Data Reported and Interpreted* and *Reliability Coefficient Cited*. This categorization resulted in seven categories for the *Data Reported and Interpreted* factor (see Table 4) and five categories for the *Reliability Coefficient Cited* factor (see Table 5).

Table 3

Standards for Reporting and Interpreting Reliability Coefficients

Standard	Source
For each total or subscore to be interpreted, estimates of reliability should be reported.	(American Educational Research Association et al., 1999, p. 31)
Reliability is a property of the scores on a test for a particular population of examinees. Authors should provide reliability coefficients of the scores for the data being analyzed.	(Wilkinson & The Task Force on Statistical Inference, 1999, p. 596)
A high α is desired, but a test need not approach a perfect scale to be interpretable.	(Cronbach, 1951, p. 331)
It would seem desirable to set .90 as a minimum reliability coefficient.	(Likert, 1932, p. 30)
Time and energy can be saved using instruments that have only modest reliability, e.g., .70.	(Nunnally & Bernstein, 1994, pp. 264-265)
Reliability should not be below .80.	(Carmines & Zeller, 1979, p. 51)
Coefficient α: Exemplary .80 or better; Extensive .70 - .79; Moderate .60-.69; Minimal <.60	(Robinson, Shaver, & Wrightsman, 1991, p. 12-18)
High reliability, test-retest consistency and internal consistency: .70 is a minimum figure.	(Kline, 1998, p. 39)
Reliability of Likert scales tends to be good . . . a reliability coefficient of .85 is often achieved.	(Oppenheim, 1992, p. 200)
A useful rule of thumb is that the reliability should be at least .70 or preferably higher.	(Fraenkel & Wallen, 2000, p. 179)
Look for high positive coefficients e.g., .60 or above	(Creswell, 2008, p. 181)

To assess the intra-rater reliability of the author's coding for the six questions, one year after the initial review and rating 10 articles published in Volume 42 (2001) were randomly selected for re-review and re-coding by the author. There was 100 % agreement between the initial coding and the re-coding made one year after the initial coding of the 10 articles. To assess inter-rater reliability of the author's coding regarding the six questions, three professors of agricultural education whose expertise includes measurement methodology were asked to review and code two randomly selected articles that had been initially reviewed and coded by the author. For each of the three independent raters there was 100 % agreement between the author's coding for the six questions and the coding made by the three professors of agricultural education.

Findings

Objective 1. Likert-scale Scores Reported and Interpreted

In 41% of the 344 articles only single-item scores were reported; however, in an additional 23% of the articles single-item scores were accompanied by a summated total score, summated subscale scores, or both (seeTable 4). Only a summated total score, summated subscale scores, or both were reported in 35% of the articles. Single-item scores were reported and interpreted in 64% of the 344 articles, whereas, summated scores – either a summated total score, summated subscale scores, or both – were reported and interpreted in 59% of the articles.

Table 4

Likert-scale Scores Reported and Interpreted

Likert-scale scores reported and interpreted	No. of articles	% of articles
Single-item scores only	142	41.3
Summated subscale scores only	77	22.4
Summated total score only	29	8.4
Single-item scores and summated subscale scores	51	14.8
Single-item scores and summated total score	21	6.1
Single-item scores and summated subscale scores and summated total score	8	2.3
Summated subscale scores and summated total score	16	4.6
Articles reporting		
Single-item scores	222	64.5
Summated subscale scores	152	44.2
Summated total score	74	21.5

Note. N = 344 articles.

Objective 2. Reliability Coefficients Cited

No reliability coefficient was cited in 30 (9%) of the articles (see Table 5). Four articles (1%) included test-retest correlation coefficients to describe the reliability of single-item mean scores and 90% of the articles cited Cronbach's α coefficients to document the reliability of a summated total score or summated subscale scores.

Table 5

Reliability Coefficients Cited for Likert-scale Scores

Reliability coefficients cited	No. of articles	% of articles
No reliability coefficient cited	30	8.7
Test-retest correlation coefficient	4	1.2
Cronbach's α: summated subscale scores	138	40.1
Cronbach's α: summated total score	131	38.1
Cronbach's α: summated subscale scores and summated total score	41	11.9
Articles citing		
No reliability coefficient	30	8.7
Test-retest correlation coefficient	4	1.2
Cronbach's α: summated subscale scores	179	52.0
Cronbach's α: summated total score	172	50.0

Note. N = 344 articles

The predominant reliability coefficient cited was a Cronbach's α coefficient, the appropriate internal consistency estimate when either a summated total score or summated subscale scores are interpreted. In contrast, only 1.2% of the articles cited a test-retest coefficient of stability, the appropriate reliability estimate when single-item mean scores are interpreted.

Objective 3. Interpretation of Likert-scale Scores

Using the Likert Score Interpretation Matrix displayed in Figure 1, the Likert-scale scores reported and interpreted in each of the 344 articles were paired with the corresponding reliability coefficients cited in each article to determine whether the interpretations of the scores were congruent or incongruent. This analysis resulted in the identification of three mutually exclusive groups of articles.

- 96 articles that exhibit **Congruent** interpretations of Likert-scale scores (see Table 6).
- 157 articles that exhibit **Incongruent** interpretations of Likert-scale scores (see Table 7).
- 91 articles that exhibit **Both Congruent and Incongruent** interpretations of Likert-scale scores (see Table 8).

	Likert-scale score interpreted		
Reliability coefficient cited	Single-item scores	Summated subscale scores	Summated total score
No coefficient cited	Incongruent	Incongruent	Incongruent
Test-retest coefficient	**Congruent**	Incongruent	Incongruent
Cronbach's α: subscale scores	Incongruent	**Congruent**	Incongruent
Cronbach's α: total score	Incongruent	Incongruent	**Congruent**

Figure 1. Likert Score Interpretation Matrix

Articles that exhibit Congruent Interpretations of Likert-scale scores. All interpretations of Likert-scale scores reported in 96 (27.9%) of the 344 articles were congruent (see Table 6). Four of the 96 articles exhibited congruent interpretations of single-item scores that were accompanied by citations of test-retest correlation coefficients documenting the reliability of the single-item scores. In 92 of the 96 articles the Likert-scale scores reported and interpreted were a summated total score, summated subscale scores, or both. Each of the summated scores interpreted was accompanied by the appropriate Cronbach's α coefficient to estimate the reliability of the summated score interpreted.

Articles that exhibit Incongruent interpretations of Likert-scale scores. All interpretations of Likert-scale scores reported in 157 (45.6%) of the 344 articles were incongruent. For 143 (91%) of the 157 articles exhibiting incongruent interpretations, single-item responses were reported and interpreted; 19 (12%) of the articles reported and interpreted summated scores, either summated subscale scores or a summated total score; and 5 (3%) of the articles reported and interpreted both a summated score and single-item scores (see Table 7).

Table 6

Articles that Exhibit ***Congruence*** *of Likert-scale Scores Interpreted and Reliability Coefficients Cited*

When scores reported are:	&	Reliability coefficients cited are:	No. of articles	% of articles	Consequence
Single-item scores only		Test-retest correlation coefficient	4	1.2	**Congruence:** Coefficient cited corresponds to scores reported.
Summated subscale scores only		Cronbach's α for subscale scores	53	15.4	**Congruence:** Coefficient cited corresponds to scores reported.
Summated total score only		Cronbach's α for total score	27	7.8	**Congruence:** Coefficient cited corresponds to scores reported.
Summated subscale scores and total summated score		Cronbach's α for summated scores and total score	12	3.5	**Congruence:** Coefficients cited correspond to scores reported.
		Total	96	27.9	

Note. N = 344 articles.

Table 7

Articles that Exhibit ***Incongruence*** *of Likert-scale Scores and Reliability Coefficients Cited*

When scores reported are:	&	Reliability coefficients cited are for:	No. of articles	% of articles	Consequence
Single-item scores only		Cronbach's α for total score	66	19.2	**Incongruence:** No reliability coefficients cited for scores reported.
Single-item scores only		Cronbach's α for subscale scores	39	11.3	**Incongruence:** No reliability coefficients cited for scores reported.
Single-item scores only		Cronbach's α for subscale scores and total score	5	1.4	**Incongruence:** No reliability coefficients cited for scores reported.
Single-item scores only		No reliability coefficient reported	28	8.1	**Incongruence:** No reliability coefficients cited for scores reported.
Single-item scores and summated total score		No reliability coefficient reported	2	0.6	**Incongruence:** No reliability coefficients cited for scores reported.
Single-item scores and summated subscale scores		Cronbach's α for total score	3	0.9	**Incongruence:** No reliability coefficients cited for scores reported.
Summated subscale scores only		Cronbach's α for total score	13	3.8	**Incongruence:** No reliability coefficients cited for scores reported.
Summated total score only		Cronbach's α for subscale scores	1	0.3	**Incongruence:** No reliability coefficients cited for scores reported.
		Total	157	45.6	

Note: N = 344 articles.

For 2 of the 19 articles reporting summated scores, incongruence resulted because no reliability coefficient was cited; for 17 of the 19 articles, incongruent interpretations resulted because summated subscale scores were accompanied by a Cronbach's α coefficient for a summated total score (16 articles) or a summated total score was accompanied by Cronbach's α coefficients for subscale scores (1 article).

For 30 (21%) of the 143 articles exhibiting incongruent interpretations of single-item responses, no reliability coefficient was cited. For 133 (79%) of the 143 articles, single-item scores were paired with a Cronbach's α coefficient of internal consistency thereby resulting in an incongruent interpretation. The Cronbach's α coefficient of internal consistency only estimates the reliability of summated (composite) scores; Cronbach's α does not estimate the reliability of single-item scores. Reporting and interpreting single-item scores derived from a Likert scale is contrary to the basic principles of Likert-scale measurement methodology. Conceptually, scores derived from Likert-type scales are composite (summated) scores. Responses to single items fail to describe adequately a construct that is defined by the multiple items comprising the Likert scale (Likert, 1932; McIver & Carmines, 1981; Nunnally & Bernstein 1994; Oppenheim, 1992). Also, the principle of aggregation empirically demonstrates that a composite score derived from the responses of multiple items is more reliable than responses to single items (Rushton, Brainerd, & Pressley, 1983; Strube 2000). The interpretation of single-item scores accompanied by a Cronbach's α coefficient is a misinterpretation of the fact that the reliability coefficient cited is a property of the *test score* – that is, the summated score derived from the Likert-type scale, not a characteristic of the Likert scale that produced the score.

Articles that exhibit Both Congruent and Incongruent interpretations of Likert-scale scores. Of the 91 articles (27.5% of 344) included in this group, 79 articles exhibited both congruent and incongruent interpretations of single-item scores, summated subscale scores, or a summated total score (see Table 8).

Each of the 91 articles exhibited congruent interpretations of summated subscale scores, a summated total score, or both because the appropriate Cronbach's α coefficient was cited for the summated scores interpreted. However, as noted in the notes for Table 8, a Cronbach's α coefficient – usually a Cronbach's α coefficient for a summated total score – was superfluously cited in 20 articles for which the corresponding summated score was not reported and interpreted.

In the 79 articles that also exhibited incongruent interpretations of Likert-scale scores, 75 of the articles reported interpretations of single-item scores. As noted for the previously described group of incongruent articles (see Table 7), incongruity resulted for this group of articles when single-item scores were interpreted. In each article reporting and interpreting single-item scores, the Cronbach's α coefficient cited pertains exclusively to the reliability of summated scores, not the single-item scores interpreted. In 4 or the 79 articles exhibiting incongruent interpretations, a summated total score was accompanied by the citation of a Cronbach's α for summated subscale scores.

Summary of findings. More than one-half (54.4%) of the 344 articles exhibited congruent interpretations of Likert-scale scores. Two-thirds (68.6%) of the articles exhibited incongruent interpretations. More than one-fifth (79 articles) exhibited both congruent and incongruent interpretations of Likert-scale scores.

Interpretations of Likert-scale scores were incongruent in 98% of the articles reporting single-item scores. When summated subscale scores, a summated total score, or both were interpreted, approximately 90% of the articles exhibited congruent interpretations.

Sixty-five percent (222 articles) of the 344 articles reporting and interpreting single-item scores reflected an analysis strategy inconsistent with the basic principles underlying Likert scale measurement methodology for quantifying social, psychological, and educational constructs. Likert-type scaling requires multiple statements (items) to define the content and meaning of the construct being quantified. The score yielded by a Likert-type scale is a composite (summation) of the responses to the multiple items comprising the scale or subscale, not responses to single items (Likert, 1932; Kerlinger, 1986; Oppenheim, 1992; Kline, 1986; Babbie, 1999). With the exception of the four articles reporting and interpreting single-item scores that cited a test-retest correlation coefficient for each item to document the reliability of single-item responses, 98% of the 222 articles reporting and interpreting single-item scores were judged to be incongruent

Table 8

Articles that Exhibit ***Both Congruent and Incongruent*** *Interpretations of Likert-scale Scores and the Reliability Coefficients Cited*

When scores reported are:	&	Reliability coefficients cited are:	No. of articles	% of articles	Consequence[a]
Single-item scores and subscale scores		Cronbach's α for subscale scores	41	11.9	**C: subscale scores** **I:** single-item scores
Single-item scores and subscale scores		Cronbach's α for subscale scores and total score[b]	7	2.0	**C: subscale scores** **I:** single-item scores
Single-item scores and total score		Cronbach's α for subscale scores[c] and total score	1	0.3	**C: total score** **I:** single-item scores
Single-item scores and total score		Cronbach's α for total score	18	5.2	**C**: **total score** **I:** single-item scores
Single-item scores and subscale scores and total scores		Cronbach's α for total score	4	1.2	**C: total score** **I:** single-item scores and subscale scores
Single-item scores and subscale scores and total score		Cronbach's α for subscale scores and total score	4	1.2	**C**: **subscale scores and total score** **I:** single-item scores
Subscale scores only		Cronbach's α for subscale scores and total score[b]	11	3.2	**C: subscale scores**
Total score only		Cronbach's α for subscale scores[c] and total score	1	0.3	**C: total score**
Subscale scores and total score		Cronbach's α for subscale scores	4	1.2	**C**: **subscale scores** **I:** total score
		Total	91	26.5	

Note. N = 344 articles.
[a]Key: **C** = Congruent; **I** = Incongruent.
[b]Cronbach's α coefficient for total score is superfluous; no total score is reported and interpreted.
[c]Cronbach's α coefficient for subscale scores is superfluous; no subscale scores are reported and interpreted.

interpretations since either no reliability coefficient was cited or a Cronbach's α coefficient was cited which pertains to summated scores, not single item scores (see Table 9).

Ninety percent of the 202 different articles reporting and interpreting a summated total score, summated subscale scores, or both were congruent interpretations. Each summated score was paired with its corresponding Cronbach's α coefficient. When a summated total score was interpreted, a Cronbach's α coefficient for the summated total score was cited; when summated subscale scores were interpreted the corresponding Cronbach's α coefficients were cited. For the 27 articles reporting summated scores that exhibited incongruent interpretations – 8 of the 27 articles exhibited both congruent and incongruent interpretations – the incongruity resulted when summated subscale scores were paired with a Cronbach's α coefficient for a summated total score, or vice versa. No reliability coefficient was cited in two articles that reported and interpreted summated scores (see Table 9).

Table 9

Summary: Articles Exhibiting Congruent and Incongruent Interpretations of Likert-scale Scores

Congruent Interpretations			**Incongruent** Interpretations		
No. articles	Likert scores reported	Reliability coefficient cited	No. articles	Likert scores reported	Reliability Coefficient cited
4	Single-item scores	Test-retest coefficient	30	Single-item scores	No coefficient cited
			188	Single-item scores	Cronbach's α coefficient
4	(1.8% of 222 articles[a])		218	(98.2% of 222 articles[a])	
116	Summated subscale scores	Cronbach's α for subscale scores	20	Summated subscale scores	Cronbach's α for total score
51	Summated total score	Cronbach's α for total score	5	Summated total score	Cronbach's α for subscale scores
16	Summated total & subscale	Cronbach's α for total & subscale	2	Summated total score	No coefficient cited
183	(90.6% of 202 articles[c])		27[b]	(13.4% of 202 articles[c]):	

Note. N = 344 articles

[a]Single-item scores were reported and interpreted in 222 of 344 articles; 142 articles reported single-item scores only; 80 articles reported both single-item scores and summated scores.

[b]Both congruent and incongruent interpretations were exhibited in 8 articles.

[c]Summated scores were reported and interpreted in 202 different articles; 152 articles reported summated subscale scores and 74 articles reported a summated total score; 24 articles reported both summated subscale scores and a summated total score.

Conclusions

For the population of 706 articles published in the *Journal of Agricultural Education* from 1995 to 2012, the Likert scale is an ubiquitous measurement methodology used to quantify psychological, social, and educational constructs.

For the target population of 344 articles published in the *Journal of Agricultural Education* from 1995 to 2012 with variables measured by a Likert-type scale, it is concluded that:

- Congruent interpretations – agreement between the Likert-scale score interpreted and the estimate of reliability cited – are exhibited in slightly more than one-half of the articles;

incongruent interpretations – dissonance between the Likert-scale score interpreted and the estimate of reliability cited – are exhibited in two-thirds of the articles.

- When summated scores are interpreted, 9 of each 10 articles exhibit congruent interpretations; when single-item scores are interpreted, fewer than 2 of each 10 articles exhibit congruent interpretations.
- The most frequently exhibited incongruity between the Likert-scale score interpreted and the reliability coefficient cited occurs when the single-item scores interpreted are accompanied by a Cronbach's α coefficient – a meaningless combination because the Cronbach's α coefficient of internal consistency is a reliability estimate exclusively applicable to summated scores.
- Reporting and interpreting single-item scores derived from a Likert-type scale violates the basic tenets of Likert-scale measurement methodology. A Likert scale requires multiple statements (items) to describe the content and meaning of the construct being quantified. The quantification of a construct requires a composite of an individual's responses to the multiple items comprising the scale, hence scores derived from Likert-type scales are summated scores. Two propositions undergird Likert scale measurement methodology: (a) conceptually, responses to single items do not describe adequately a construct that is defined by the content of multiple items, and (b) the empirically verifiable principle of aggregation demonstrates that a composite score derived from responses to multiple items is more reliable than are responses to single items (Likert, 1932; McIver & Carmines, 1981; Nunnally & Bernstein, 1994; Oppenheim, 1992).

Recommendations

It is recommended that university faculties of agricultural and extension education, communication, and leadership examine graduate courses and continuing education seminars and workshops to insure that authors and potential authors of research published in the *Journal of Agricultural Education* receive instruction regarding the concepts of reliability of measurement and the theory and practice of Likert-scale measurement methodology.

It is recommended that referees of manuscripts submitted to the *Journal of Agricultural Education* be more diligent in reviewing manuscripts regarding the calculation and analysis of test scores derived from Likert scales and the reliability coefficients cited to document the reliability of Likert scores reported and interpreted (Roberts, et al., 2011).

The following strategies and procedures for reporting and interpreting Likert-scale scores are recommended[2].

- ✓ Construct a frequency table stating each item comprising the Likert scale that records the percentage of respondents choosing each option of the response continuum. Use the content of the multiple items comprising the scale and any subscales within the scale to describe the meaning and *name* for the constructs.
- ✓ For *each respondent*, calculate a summated total score for the multiple items comprising the scale and summated subscale scores for any subscales within the Likert scale.
- ✓ Compute and report the appropriate Cronbach's α coefficient for the summated total score and any summated subscale scores. Specifically identify the Cronbach's α coefficients that correspond to the summated total score and any summated subscale scores.
- ✓ For the *group of respondents*, compute descriptive statistics for the summated total score and any summated subscale scores, including central tendency (mean, median, and mode), variability (standard deviation and range), skewness, and symmetry (kurtosis).

[2] An appendix to the article is available from the author that illustrates, using an actual database, the scoring, reporting, analysis, and interpretation strategies recommended.

- ✓ For the summated total score and any summated subscale scores, construct a frequency table or histogram and record as a part of the table or chart the descriptive statistics for central tendency, variability, and skewness.
- ✓ In the text accompanying each frequency table or histogram, describe the content and meaning of the construct and any subconstructs for the summated total score and summated subscale scores that are reported and interpreted.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

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

Babbie, E. (1999). *The basics of social research.* New York, NY: Wadsworth.

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

Creswell, J. W. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* Upper Saddle River, NJ: Pearson.

Cronbach, L. J. (1951). Coefficient alpha and internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Cronbach, L. J. (1984). *Essentials of psychological testing.* New York, NY: Harper & Row.

Edwards, A. L. (1957). *Techniques of attitude scale construction.* New York, NY: Appleton-Century-Crofts.

Fraenkel, J. R., & Wallen, N. E. (2000). *How to design and evaluae research in education.* New York, NY: McGraw-Hill.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis.* Upper Saddle River, NJ: Prentice Hall.

Isaac, S., & Michall, W. B. (1995). *Handbook in research and evaluation.* San Diego, CA: Educational and Industrial Testing Service.

Kerlinger, F. N. (1986). *Foundations of behavioral research.* New York, NY: Holt, Rinehart and Winston.

Kline, P. (1998). *The new psychometrics: Science, psychology and measurement.* New York, NY: Routledge.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 1-55.

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

Nunnally, J. C., & Berstein, I. H., (1994). *Psychometric theory.* New York, NY: McGraw-Hill.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing, and attitude measurement.* New York, NY: Printer Publishers.

Roberts, T. G., Barrick, R. K., Dooley, K. E., Kelsey, K. D., Raven, M. R., & Wingenbach, G. J. (2011). Enhancing the quality of manuscripts submitted to the *Journal of Agricultural Education*: Perceptions of experienced reviewers. *Journal of Agricultural Education*, *52*(3), 1-5.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). Criteria for scale selection and evaluation. In J. P. Robinson & L. S. Wrightsman (Eds.). *Measures of personality and psychological attitudes* (pp. 1-16). New York, NY: Academic Press.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, 94(1), 18-38.

Spector, P. E. (1992). *Summated rating scale construction: An introduction.* Newbury Park, CA: Sage.

Strube, M. J. (2000). Reliability and generalizability theory. In L. G. Grimm and P. R. Yarnold (Eds.). *Reading and understanding more multivariate statistics* (pp. 23-66). Washington, DC: American Psychological Association.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594-604.