Journal of Rural Social Sciences, 31(3), 2016, pp. 126–139.
Copyright © by the Southern Rural Sociological Association

ANOTHER LOOK AT LIKERT SCALES*

FERN K. WILLITS
THE PENNSYLVANIA STATE UNIVERSITY

GENE L. THEODORI
SAM HOUSTON STATE UNIVERSITY

AND

A.E. LULOFF
THE PENNSYLVANIA STATE UNIVERSITY

ABSTRACT

Perhaps the most widely used means for assessing survey respondents' personal attitudes consists of a series of stem-statements followed by an odd or even number of ordered, bipolar-named categories. Such statements, known as Likert items, are named for Rensis Likert whose classic studies of attitude measurement were first published in 1932. Almost from the beginning, methodologists and psychometric scholars have raised questions concerning the number of items deemed necessary to form an attitude scale, the number and meaning of various answer categories, and the appropriate statistical methods to use in analyzing the resulting data. These deliberations are summarized. We conclude that, while continuing research on the meaning and uses of Likert scales is needed, many criticisms directed against their usage are unwarranted. Both Likert items and multi-item scales represent useful means for researchers seeking information on subjects' attitudes.

Social scientists have produced an extensive literature on the nature and social correlates of public attitudes. The findings from some of these studies may both aid in social decision-making in a democratic society and contribute to our understanding of the sources of human behavior. Early scholars assumed attitudes were not an acceptable area of scientific inquiry since attitudes cannot be observed directly and therefore need to be inferred or deduced from individuals' actions. Thurstone (1928) challenged this position in his paper entitled "Attitudes Can Be Measured."

Although the methods suggested by Thurstone were cumbersome and are seldom used today (Thurstone and Chave 1929), his work was quickly followed by that of others. One method, developed in 1932 as a doctoral dissertation in psychology at Columbia University, has come to dominate current attitude

*An earlier version of this paper was presented at the annual meeting of the Rural Sociological Society, Madison, WI, August 2015.

measurement (Likert 1932). This approach consisted of asking subjects to indicate the extent to which they agreed or disagreed with each of a series of statements related to the focus of the desired attitude. The resulting answers were then scored and summed to yield a composite value used to index the respondents' attitudes toward the topic of interest. Although the developer of this simple, pragmatic method for measuring attitudes went on to have a distinguished academic career as a renowned survey statistician, cofounder and director of the Survey Research Center at the University of Michigan, president of the American Statistical Association, and leader in cutting-edge work on participative business management practices (Seashore 1982), he is most often identified today for his development of this attitude scaling procedure. His name was Rensis Likert ("lick-urt") and his methods are called "Likert Scaling."

CHARACTERISTICS OF LIKERT SCALES

"Likert Scales" consist of a series of related "Likert-type items" – statements concerning a specific referent, namely the focus of the attitude to be measured (Desselle 2005; Likert 1932). A balance of both positive and negative items is generally recommended to reduce response-set bias. Subjects indicate their feelings concerning each item on a bipolar scale such as "strongly agree, agree, undecided, disagree, and strongly disagree." Responses for each subject are scored from one (1) to five (5), with negative items reverse-coded. The scores for the individual items are then summed to obtain a Summated Rating Score or Likert Scale value for each respondent. Alternatively, the mean scores of the responses of each subject can be used so that the scale scores fall in the same 1 to 5 range as the individual items. Although these five category response alternatives are common, three, four, six, seven, and more have also been used. Factor analysis (Flora and Curran, 2004) and/or item analysis, including item-to-item, item-to-total correlations and/or reliability measures such as Cronbach's Alpha (Cronbach 1951), may be used to assess the extent to which the separate items are assessing a single attitude dimension.

Although Likert-type items and Likert Scales have been widely adopted throughout the social science research communities, the method is not without controversy. Through the years, the procedures involved in their derivation and use have been the subjects of debate by social science methodologists, psychometric scholars, and applied researchers concerning the number of necessary items, the number, and nature of the response categories, and the uses of the summated and item scores. The purpose of this paper is to explore the meanings and implications

of these various issues and, by doing so, contribute to ongoing dialogue in this area. Specifically, the following three issues are addressed:

1) How many Likert-type items are needed for a Likert attitude scale?
2) What and how many response categories should be presented?
3) What are the meanings of the obtained responses? How can they be analyzed?

HOW MANY LIKERT-TYPE ITEMS ARE NEEDED FOR A LIKERT SCALE?

Attitudes are latent variables. Although they influence behavior, attitudes cannot be directly observed; they must be inferred through a person's various actions or pronouncements. Likert Scales seek information for understanding a subject's attitude by combining the individual's responses to a series of opinion questions designed to address relevant aspects of the attitude in question. The Likert Summated Rating Scale procedure outlined above is one such measure. The use of multiple items, rather than a single question, is expected to yield an index that is more reliable, valid, and discriminatory than a single item. Single items have considerable random measurement error. Such variation is expected to average out when multiple indicators are used. That is, a total scale developed from multiple items is expected to be more consistent and reliable than responses to any single item (Carmines and Zeller 1979; McIver and Carmines 1981; Nunnally and Bernstein 1994; Spector 1992). Further, single items lack scope and often fail to validly measure the total meaning of the concept: "It is very unlikely that a single item can fully represent a complex theoretical concept or any specific attribute for that matter" (McIver and Carmines 1981:151).

Determining the appropriate number of items to include in a Likert Scale remains problematic. For many, "more" is often seen as "better," since using many items allows for capturing the nuances of complex attitude structures while diluting the impact of random variation in single items. Indeed, Nunnally (1978:243) noted: "other things being equal, a long test is a good test." Although no fixed rules exist concerning the number of items to include in the final scale, at least four are needed for evaluation of internal consistency (Diamantopoulos et al. 2012). Moreover, while reliability measures increase as the number of items increases above five, each addition makes progressively less impact on scale reliability (Carmines and Zeller 1979; Hinkin 1995). As a result, from a practical standpoint, approximately five, six, or seven items have been suggested as adequate for most constructs (Hinkin 1998). Still, these numbers refer to the final scale. Typically, some items developed for a

given study fail to meet inclusion criteria suggesting that as many as twice the desired number might need to be included in an initial survey.

Not surprisingly, much emphasis has been given to the importance of using multi-item scale scores to measure attitudes and other social and psychological constructs (Churchill 1979; Maranell 1974; Oskamp and Schultz 2005; Peter 1979). Indeed, Gliem and Gliem (2003:82, 88) bluntly stated:

> It is not appropriate to make inferences based upon the analysis of single-item questions which are used in measuring a construct ... analysis of data 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.

Despite these criticisms, the use of individual attitude items has an important place in social science research as scholars seek to understand the views of residents concerning various social issues, to predict the values of consumers, or to assess the political views of constituents. Although none would argue that complex psychological issues are captured by a single item (Loo 2002), when a construct is narrow in scope, unidimensional, concrete, and unambiguous a single item is often as useful as more complex multi-item scales in predictive validity (Bergkvist and Rossiter 2007; de Boer et al. 2004; Diamantopoulos et al. 2012; Sackett and Larson 1990; Rossiter 2002; Wanous, Reichers, and Hudy 1997).

Moreover, many researchers defend the use of data from individual Likert-type items (Clason and Dormody 1994) arguing that single item responses may often be the "best that can be obtained" in practice. The inclusion of batteries of attitude items not only increases the length and cost of data collection, but also contributes to greater respondent burden and fatigue and may lead to higher refusal rates (Dillman, Smyth, and Christian 2014; Drolet and Morrison 2001). These issues are particularly relevant when essentially redundant items are included to increase scale reliability coefficients.

Single items may also be useful when asking respondents to provide an overall evaluation of more complex phenomena. Thus, for example, a general question concerning subjects' overall satisfaction with their communities (although clearly multidimensional) may be more relevant than simply summing their expressed satisfaction with various facets of community life (e.g., public services, neighborliness, economic opportunities, etc.). In the former case, the respondent has

the opportunity to weight the parts subjectively; in the latter, the researcher provides the weightings, generally by assuming the items are equal in importance.

WHAT RESPONSE CATEGORIES SHOULD BE PRESENTED FOR EACH ITEM?

In Likert's initial presentation, subjects were asked to respond to each item on a five-category bipolar scale by indicating whether they "strongly approved, approved, neither approved nor disapproved, disapproved, or strongly disapproved" to each of the positive or negative opinion statements provided (Likert 1932). The psychometric model for such items assumes the responses represent a single continuous latent construct with opposite feelings expressed at the endpoints. Thus, "strongly opposed" was taken as the direct opposite of "strongly support" with the middle category representing a position midway on that continuum. Although he used five response categories, Likert was clear in indicating that both more and fewer numbers of alternatives were also appropriate. Nevertheless, the most common format used today employs the five categories of "strongly agree, agree, undecided (or neither agree nor disagree), disagree, and strongly disagree." The use of such named categories is user-friendly and has been found to provide acceptable levels of reliability (Dillman et al. 2014:159). The responses are then scored from 1 to 5 (or 5 to 1) for each item thus assuming the intervals between responses are equal. Tradition, ease of use, and comparability with other studies, both currently and historically, support the utility of using this five-category response pattern. Despite survey mode (paper and pencil, online, telephone, or face to-face), these categories are easily presented and convey the idea of a continuum of responses.

However, various alternative formats have also been used. Extending the number of categories allows for greater differentiation in responses. A seven-category response scale (very strongly agree, strongly agree, agree, undecided, disagree, strongly disagree, very strongly disagree) is straightforward and allows for greater differentiation in responses. However, using more than seven similarly named categories (e.g., very, very strongly agree; very strongly agree; strongly agree, agree, etc.) is awkward and confusing. As an alternative, and to reinforce the numerical nature of the resulting score, labeled endpoints connected by a horizontal line with equally spaced unlabeled or numbered gradients can be used to encourage subjects to visualize and record their responses on a numerical scale. Doing so also allows for a straightforward extension to the use of more categories and increased sensitivity of the measuring instrument. Gathering such detailed information could

be especially important if a distribution of scores is heavily skewed with many subjects falling on one side of the mid-value (Cummins and Gullone 2000).

Researchers assessing the *relative* reliability of various formats have reported mixed findings. Some researchers have found little relationship between the number of alternative responses presented and validity or reliability indicators (Jacoby and Matell 1971; Matell and Jacoby 1971; Schutz and Rucker 1975). Others have suggested that as the number of responses increased from approximately five to 10, reliability measures increased (Green and Rao 1970; Preston and Colman 2000). Still others have reported that coefficient alpha reliabilities increased up to the use of five points, but then leveled off with increasing number of response categories (Lissitz and Green 1975). Perhaps the variation in these findings resulted from differences in subject characteristics and/or topics addressed by the questions. Whatever the reason, previous studies focusing on reliability issues fail to provide clear guidance concerning the most appropriate number of response categories to use in Likert-type items.

Critical to the choice of how many categories are to be used are questions concerning the capability and willingness of respondents to make the detailed distinctions requested of them. Although subjects are clearly able to distinguish between "agreeing" and "strongly agreeing," expecting them to report four or five differing levels of agreeing or disagreeing may not be reasonable (Cox 1980). When respondents were asked to choose among their preferences regarding the desirability of using varying numbers of alternative formats, most chose more than four but less than 11 categories (Preston and Colman 2000).

The inclusion of an "undecided" or "neither agree nor disagree" response scored between "agree" and "disagree" was part of Likert's original formulation and continues to be used by most researchers. However, the meaning of this middle category is ambiguous. Does it imply the subject: 1) has no opinion; (2) has a "balanced" view in terms of evaluation; (3) is indifferent/does not care; and/or (4) does not understand the question? (Dubois and Burns 1975; Kulas and Stachowski 2009; Shaw and Wright 1967; Tourangeau, Smith, and Rasinski 1997). No matter why a respondent chooses such a middle category, he or she has been unable or unwilling to state an opinion. Thus, it seems inappropriate to score it as quantitatively halfway between "agree" and "disagree" but rather to define this category as *qualitatively* different from its adjacent categories by treating it as a separate dichotomous variable, which may present important information worthy of analysis. Thus, some previous analysis has found that the social characteristics of respondents who state an opinion differ from those who agree or disagree with

the issue in question (Krosnick 1999; Willits, Theodori, and Luloff 2016). The use of an alternative format in which respondents indicated whether they had an opinion (yes/no), followed by: If "yes," do you "strongly agree," "agree," "disagree," or "strongly disagree" with the item. Regardless, the inclusion of some means of differentiating those who cannot or will not state an opinion is an important component of any Likert analysis. Additional research on differing samples and focusing on other topics is needed to extend our understanding of the meaning of the middle category.

WHAT ARE THE MEANINGS OF THE OBTAINED RESPONSES? HOW CAN THEY BE USED IN ANALYSIS?

Despite the widespread use of Likert Scales and Likert-type items in social science research, there is considerable controversy among scholars concerning the meaning and uses of the obtained data: Given the nature of Likert-type items and Likert scales, what statistical/analytical procedures are appropriate?

Writing more than 60 years ago, S.S. Stevens (1946) described a hierarchy of measurement that consisted of: nominal scales (measurement by categories without numerical representation); ordinal scales (measurement by ranking or ordering of categories or items without information concerning the distances between them); interval scales (measurement using a unit of measure with ordering and distance indicators); and ratio scales (interval-type scales with an absolute zero value). Stevens (1968) went on to argue that these types of measurement defined what statistical procedures were permissible for analytic purposes. For Stevens, means, standard deviations, t-tests, product moment correlations, and analysis of variance were seen as permissible only for analyzing variables measured by interval or ratio scales; ordinal scaled variables were appropriately analyzed using statistical methods dealing with ranks, including medians, ranges, rank correlations, and other nonparametric tools. The inclusion of these pronouncements in several popular statistics textbooks (e.g., Blalock 1960; Siegel 1956), and, more recently, in at least one major computer package (SPSS 2008) have fostered widespread acceptance of these caveats. Following Stevens' dictates, both Likert scales and Likert-type items (1) constitute ordinal (not interval) scales; and, (2) fail to meet the statistical assumptions of normality and homoscedasticity, thus ruling out the use of standard parametric statistical tools.

Stevens' assertion (1946) that the method of measurement (nominal, ordinal, interval, or ratio) proscribes the types of statistical operations that are appropriate has been challenged by many methodologists, mathematicians, and statisticians

through the years (Boneau 1960; Borgatta and Bohrnstedt 1980; Gaito 1980; Labovitz 1970; Velleman and Wilkinson 1993). At issue is the extent to which the validity of the statistical results are affected by the application of common parametric tests to data that do not meet their mathematical assumptions, including normality and homoscedasticity. Statisticians and methodologists have explored the issues concerning the impact of such assumption violations on the findings from a wide range of empirical and theoretical analyses. These studies have consistently documented the robustness of the resulting analysis – the likelihood that the tests will give appropriate conclusions even when their mathematical assumptions are violated (Baker et al. 1966; Carifio and Perla 2007; Norman 2010).

The idea that Likert *scales* which combine the summated effects of multiple Likert-type *items* has become widely accepted as resulting in quantitative interval scale scores (Allen and Seaman 2007; Boone and Boone 2012; Brown 2011; Carifio and Perla 2007; Clason and Dormody 1994). However, it can be argued that responses to Likert type *items* can also be treated as interval scales. Likert suggested the ordered named responses to an item imply an underlying continuum or quantitative score to respondents. The similarities described above among responses obtained using five, seven, nine or more named categories and those obtained from unlabeled ruler marks on a visual scale support this expectation and suggest responses to single items with five or more item responses represent measures that can be appropriately viewed as interval in nature. In these terms, stand-alone items and those included in multi-item scales score can be analyzed separately to provide information on subjects' responses to specific aspects or components of the whole of which they are a part.

Indeed, if one is unwilling to accept the idea that individual item responses represent numerical (interval) ratings, seeing how combining these responses into composite scores (sums or averages) magically converts them into interval scales is difficult. The use of composite scores based on multiple items can provide more stable (less random fluctuation) ratings and measures of more complex phenomena than can individual item responses. However, responses to individual items also represent information that is no less numerical.

Norman (2010:631) summarized the matter succinctly:

> Parametric statistics can be used with Likert data, with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of 'coming to the wrong conclusion'. These findings are consistent with

empirical literature dating back nearly 80 years. The controversy can cease (but likely won't).

SUMMARY

The widespread use of Likert-type items and multi-item Likert Scales has led to a host of myths and misunderstandings concerning their appropriate use as discussed above. The ideas presented here are not new. Increasingly scholars are reexamining the rules of scientific inquiry. In this paper, we have, for the most part, refrained from suggesting best practices that constrain or restrict continuing exploration of multiple procedures and perspective. It is from such dictates that myths can arise and become dogma.

Myth 1: *Likert Scales MUST contain multiple items; a single Likert-type item used to measure a concept is useless.* For complex concepts (e.g., environmental attitudes; community satisfaction, intelligence, political conservatism, etc.), multi-item scales may be needed to provide a global summary of a respondent's views about these topics and be less subject to random fluctuation than are individual items. However, single items are appropriate when the referenced concept is singular, concrete, and understandable to the respondent. Single items can also be useful in obtaining variation in respondents' subjective evaluations of more complex phenomena by providing information on differentiation among specific components that make up a generalized whole.

Myth 2: *The "middle category" (i.e., "undecided" or "neither agree nor disagree") can be deleted from the Likert item response categories since it provides no useful data for analysis.* "Undecided" is a valid response in situations where subjects may have no knowledge or no opinion. To not offer such a middle category forces subjects to omit the item or provide an incorrect response. Analysis to determine the characteristics of these undecided respondents may be useful in clarifying the meaning of such answers and in adding to our understanding of human behavior.

Myth 3: *Although Likert Scales based on multiple items yield numerical (intervally scaled) data, information from single Likert items must be treated as ordinal data.* Although single Likert items allow for only a few (often five) named responses, subjects often view these as points on a continuum from low to high with response distributions similar to those obtained on a scaled line with equal intervals between points on that scale.

Myth 4: *Data from Likert Scales and Likert items cannot be analyzed using statistical tests such as t-tests, Analysis of Variance, and Pearsonian Correlations because the parametric assumptions of these tests are not met.* These parametric tests are very robust

– meaning they maintain their essential validity even when their assumptions (normality, homogeneity of variances, sample size. and interval measurement) are not strictly met.

CONCLUSION

The use of Likert Scales and Likert–type items have served the research community well though the years. Despite criticisms leveled against their use, analysis using Likert scales and Likert type items has contributed to advancements of knowledge in sociology, psychology, political science, biology, economics, marketing, medicine, and other fields. This paper has sought to summarize the current state of knowledge and practice in this area. It has shown that many issues raised by critics are essentially myths. Yet they are repeatedly quoted as reasons for critiquing and rejecting others creative research findings. We would agree with Norman (2010) and Bacchetti (2002) that strong evidence of a review culture in science encourages criticism for its own sake, often focusing on inappropriate statistical and methodological dogmatism and rules of presumed legality. Such an orientation does little to advance the goals of scientific inquiry. Research is needed that explores the empirical effects, substantive meanings, and practical implications of our findings. In doing so, recognizing that sometimes the perfect can be the enemy of the good is important, and:

> … learning is a sequential process and that theory [and methodology] sometimes fall short of providing the sort of prior information one would like. Occasional sinning, therefore, may be inevitable (Wallace 1977: 443).

REFERENCES

Allen, Isabel E. and Christopher A. Seaman. 2007. "Likert Scales and Data Analyses." *Quality Progress* 40(7):64–5.

Bacchetti, Peter. 2002. "Peer Review of Statistics in Medical Research: The Other Problem." *British Medical Journal* 324(7348):1271–3.

Baker, Bela. O., Curtis D. Hardyck and Lewis F. Petrinovich. 1966. "Weak Measurements vs. Strong Statistics: An Empirical Critique of S. S. Stevens' Proscriptions on Statistics." *Educational and Psychological Measurement* 26(2):291–309.

Bergkvist, Lars and John R. Rossiter. 2007. "The Predictive Validity of Multiple-Item Versus Single-Item Measures of the Same Constructs." *Journal of Marketing Research* 44(2):175–84.

Blalock, Hubert M. 1960. *Social Statistics.* New York: McGraw-Hill.

Boneau, C. Alan. 1960. "The Effects of Violations of Assumptions Underlying the T Test." *Psychological Bulletin* 57(1):49–64.

Boone, Harry N. and Deborah A. Boone. 2012. "Analyzing Likert Data." *Journal of Extension* 50(2). Retrieved October 7, 2015 (http://www.joe.org/joe/2012april/tt2.php).

Borgatta, Edgar F. and George W. Bohrnstedt. 1980. "Level of Measurement: Once Over Again. *Sociological Methods & Research* 9(2):147–60.

Brown, James D. 2011. "Likert Items and Scales of Measurement." *SHIKEN: JALT Testing and Evaluation SIG Newsletter* 15(1):10–4.

Carifio, James and Rocco J. Perla. 2007. "Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends About Likert Scales and Likert Response Formats and Their Antidotes." *Journal of Social Sciences* 3(3):106–16.

Carmines, Edward G. and Richard A. Zeller. 1979. *Reliability and Validity Assessment.* Thousand Oaks, CA: Sage.

Churchill, Gilbert A. 1979. "A Paradigm for Developing Better Measures of Marketing Constructs." *Journal of Marketing Research* 16(1):64–73.

Clason, Dennis L. and Thomas J. Dormody. 1994. "Analyzing Data Measured by Individual Likert-Type Items." *Journal of Agricultural Education* 35(4):31–5.

Cox III, Eli P. 1980. "The Optimal Number of Response Alternatives for a Scale: A Review." *Journal of Marketing Research* 17(4):407–22.

Cronbach, Lee J. 1951. "Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16(3): 297–334.

Cummins, Robert A. and Eleonora Gullone. 2000. "Why We Should Not Use 5-Point Likert Scales: The Case for Subjective Quality of Life Measurement." Pp. 74–93 in *Proceedings, Second International Conference on Quality of Life in Cities.* Singapore: National University of Singapore.

de Boer, A.G.E.M., J.J.B. van Lanschot, P.F.M. Stalmeier, J.W. van Sandick, J.B.F. Hulscher, J.C.J.M. de Haes, and M.A.G. Sprangers. 2004. "Is a Single-item Visual Analogue Scale as Valid, Reliable and Responsive as Multi-item Scales in Measuring Quality of Life?" *Quality of Life Research* 13(2):311–20.

Desselle, Shane P. 2005. "Construction, Implementation, and Analysis of Summated Rating Attitude Scales." *American Journal of Pharmaceutical Education* 69(5):1–11.

Diamantopoulos, Adamantios, Marko Sarstedt, Christoph Fuchs, Petra Wilczynski and Sebastian Kaiser. 2012. "Guidelines for Choosing between Multi-Item and

Single-Item Scales for Construct Measurement: A Predictive Validity Perspective." *Journal of the Academy of Marketing Science* 40(3):434–49.

Dillman, Don A., Jolene. D. Smyth, and Leah M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method (4th Edition)*. Hoboken, NJ: John Wiley.

Drolet, Aimee L. and Donald G. Morrison. 2001. "Do We Really Need Multiple-item Measures in Service Research? *Journal of Service Research* 3(3):196–204.

DuBois, Bernard and John A. Burns. 1975. "An Analysis of the Meaning of the Question Mark Response Category in Attitude Scales." *Educational and Psychological Measurement* 35(4):869–84.

Flora, David B. and Patrick J. Curran. 2004. "An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis with Ordinal Data." *Psychological Methods* 9 (4): 466.

Gaito, John. 1980. "Measurement Scales and Statistics: Resurgence of an Old Misconception." *Psychological Bulletin* 87 (3):564–7.

Gliem, Joseph A.and Rosemary R. Gliem. 2003. "Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-type Scales." Paper presented at Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, Columbus, OH. Retrieved October 7, 2015 (http://www.ssnpstudents.com/wp/wp-content/uploads/2015/02/Gliem-Gliem.pdf).

Green, Paul E. and Vithala R. Rao. 1970. "Rating Scales and Information Recovery. How Many Scales and Response Categories to Use?" *Journal of Marketing* 34(3):33–9.

Hinkin, Timothy R. 1995. "A Review of Scale Development Practices in the Study of Organizations." *Journal of Management* 21(5):967–88.

_____. 1998. "A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires." *Organizational Research Methods* 1(1):104–21.

Jacoby, Jacob and Michael S. Matell. 1971. "Three-Point Likert Scales are Good Enough." *Journal of Marketing Research* 8(4):495–500.

Krosnick, Jon A. 1999. "Survey Research". *Annual Review of Psychology*, 50(1):537–67.

Kulas, John T. Alicia A. Stachowski. 2009. "Middle Category Endorsement in Odd-Numbered Likert Response Scales: Associated Item Characteristics, Cognitive Demands, and Preferred Meanings." *Journal of Research in Personality* 43(4):489–93.

Labovitz, Sanford. 1970. "The Assignment of Numbers to Rank Order Categories." *American Sociological Review* 35(3):515–24.

Likert, Rensis. 1932. "A Technique for the Measurement of Attitudes." *Archives of Psychology* 22:5–55.

Lissitz, Robert W. and Samuel B. Green. 1975. "Effect of the Number of Scale Points on Reliability: A Monte Carlo Approach." *Journal of Applied Psychology* 60(1):10–3.

Loo, Robert. 2002. "A Caveat on Using Single-item Versus Multiple-item Scales." *Journal of Managerial Psychology* 17(1):68–75.

Maranell, Gary M. 1974. *Scaling: A Sourcebook for Behavioral Scientists*. Chicago, IL: AldineTransaction.

Matell, Michael S. and Jacob Jacoby. 1971. "Is There an Optimal Number of Alternatives for Likert Scale Items?: Study I: Reliability and Validity." *Educational and Psychological Measurement* 31:657–74.

McIver, John P. and Edward G. Carmines. 1981. *Unidimensional Scaling*. Newbury Park, CA: Sage Publications.

Norman, Geoff. 2010. "Likert Scales, Levels of Measurement and the "Laws" of Statistics." *Advances in Health Sciences Education* 15(5):625–32.

Nunnally, Jum C. 1978. *Psychometric Theory*. New York: McGraw Hill.

Nunnally, Jum C. and Ira H. Bernstein. 1994. *Psychometric Theory*, 3rd edition. New York: McGraw Hill.

Oskamp, Stuart and P. Wesley Schultz. 2005. *Attitudes and Opinions*. New York: Psychology Press.

Peter, J. Paul. 1979. "Reliability: A Review of Psychometric Basics and Recent Marketing Practices." *Journal of Marketing Research* 16(1):6–17.

Preston, Carolyn C. and Andrew M. Colman. 2000. "Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences." *Acta Psychologica* 104(1):1–15.

Rossiter, John R. 2002. "The C-OAR-SE Procedure for Scale Development in Marketing." *International Journal of Research in Marketing* 19(4):305–35.

Sackett, Paul R. and James R. Larson, Jr. 1990. "Research Strategies and Tactics in Industrial and Organizational Psychology." Pp. 419–89 in *Handbook of Industrial and Organizational Psychology, Vol. 1* (2nd edition), M.D. Dunnette & L.M. Hough (eds). Palo Alto, CA: Consulting Psychologists Press.

Seashore, Stanley E. 1982. "Obituary: Rensis Likert (1903-1981)." *American Psychologist* 37(7):851–3.

Schutz, Howard G. and Margaret H. Rucker. 1975. "A Comparison of Variable Configurations Across Scale Lengths: An Empirical Study." *Educational and Psychological Measurement* 35(2):319–24.

Shaw, Marvin E. and Jack Mason Wright. 1967. *Scales for the Measurement of Attitudes*. New York: McGraw Hill.

Siegel, Sidney. 1956. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw Hill.

Spector, Paul E. 1992. *Summated Rating Scale Construction: An Introduction*. Newbury Park, CA: Sage.

SPSS Inc. 2008. *SPSS Statistics for Windows, Version 17.0*. Chicago, IL: SPSS Inc.

Stevens, S.S. 1946. "On the Theory of Scales of Measurement." *Science* 103(2684):677–80.

_____. 1968. "Measurement, Statistics, and the Schemapiric View." *Science* 161(3844):849–56.

Thurstone, L. L. 1928. "Attitudes Can Be Measured." *American Journal of Sociology* 33(4):529–54.

Thurstone, L. L. and E.J. Chave. 1929. *The Measurement of Attitude*. Oxford, England: University of Chicago Press.

Tourangeau, Roger, Tom W. Smith, and Kenneth A. Rasinski. 1997. "Motivation to Report Sensitive Behaviors on Surveys: Evidence from a Bogus Pipeline Experiment." *Journal of Applied Social Psychology* 27(3):209–22.

Velleman, Paul F. and Leland Wilkinson. 1993. "Nominal, Ordinal, Interval, and Ratio Typologies are Misleading". *The American Statistician* 47(1):65–72.

Wallace, T. Dudley. 1977. "Pretest Estimation in Regression: A Survey." *American Journal of Agricultural Economics* 59(3):431–43.

Wanous, John P., Arnon E. Reichers, and Michael J. Hudy. 1997. "Overall Job Satisfaction: How Good are Single-item Measures?" *Journal of Applied Psychology* 82(2): 247–52.

Willits, Fern K., Gene L. Theodori, and A.E. Luloff. 2016. "Self-Reported Familiarity of Hydraulic Fracturing and Support for Natural Gas Drilling: Substantive and Methodological Considerations." *Journal of Rural Social Sciences* 31(1):83–101.