The Likert Scale
What It Is and How To Use It

Maj Katherine A. Batterton, USAF, katherine.batterton@us.af.mil
Capt Kimberly N. Hale, USAF, kimberly.hale.2@us.af.mil

Introduction

Surveys are ubiquitous analytic tools used throughout the military services. From assessing an organization's health to soliciting opinions on policies, surveys are often used to gather data. In general, these surveys focus on measuring attitudes and these attitudes are almost universally measured using Likert scales. The results of these surveys often drive policy decisions. So, it is critical to develop surveys carefully and conduct appropriate analysis to provide accurate insights.

Unfortunately, military analysts are not typically trained in survey development or analysis. The initial skills training for Air Force and Army operations research analysts does not include a formal course on survey analysis or related statistical methods (Henry and Smith 2015). Lack of formal training is made more problematic with the large body of literature that cannot agree on the appropriate treatment of Likert scales. In fact, a quick Google search of "Likert scales analysis" returns competing methodological opinions. Some claim that Likert scales are ordinal, and only nonparametric statistical methods are appropriate (Allen and Seaman 2007). Others, including the Director, Operational Test and Evaluation for the Office of the Secretary of Defense, advocate using parametric methods and treating Likert scales as continuous (Office of the Secretary of Defense 2015, Norman 2010). While methods for analyzing survey data are not new and many papers attempt to highlight the proper treatment of Likert scales, it is clear confusion on Likert scales remains (Carifio and Perla 2008, Boone and Boone 2012).

To shine some light on these issues, we will clarify what actually constitutes a Likert scale, describe a method used to create them, and demonstrate analysis of a Likert scale with an example. We end with a brief discussion of important and often overlooked considerations for survey development that create a solid foundation for analyses.

The Likert Scale: What It Is (and What It Isn't)

In 1932, Rensis Likert published the paper "A Technique for the Measurement of Attitudes," which introduced a method for measuring attitudes and his eventual namesake, the Likert scale. Since then, lackadaisical use of terminology has generated a heated debate in the literature on the best statistical methods for analyzing Likert scales (Jamieson 2004, 2005; Pell 2005; Carifio and Perla 2007). The primary disagreement centers on the treatment of the data as ordinal or interval. As it turns out, much of the disagreement stems from confusion and could be cleared up by simply clarifying the definition of a Likert scale.

In his 1932 paper, Likert described a simple and powerful method to construct an attitude scale, built on the premise that groups of related questions measure a subject's attitude about some issue addressed by those questions. He demonstrated that summing the responses from the related questions resulted in a useful and tractable measure of the underlying attitude (Likert 1932).

Likert's "Survey of Opinions" used several different response formats.[a] One of these response formats offered five possible choices: strongly disagree, disagree, undecided, agree, and strongly agree. This type of response format is commonly misidentified as a Likert scale. An individual response on a question is a Likert-type response (Figure 1). Likert-type responses can be thought of as ordered, but since the distance between each response choice is not necessarily constant or well-defined, Likert-type responses

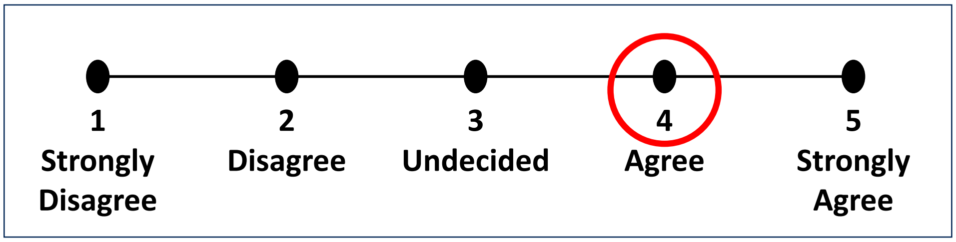

Figure 1. Illustration of a Likert-type response.

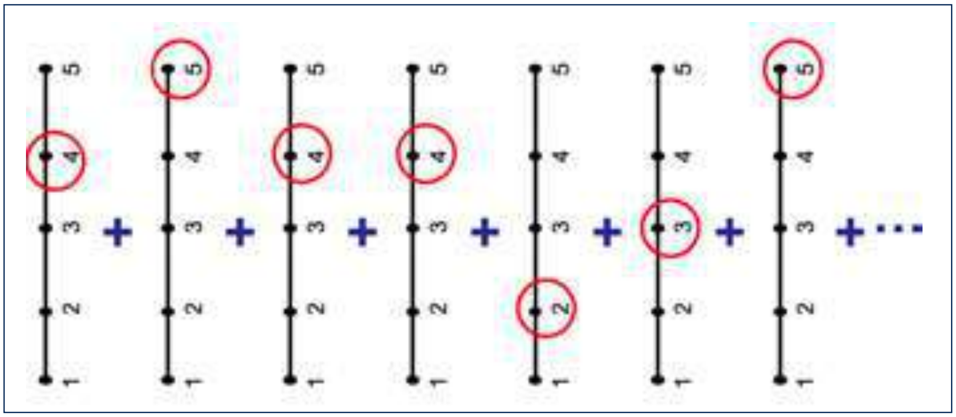

Figure 2. Illustration of a Likert scale constructed as the sum of Likert-type responses from related questions.

reflect ordinal rather than interval data. This data format makes the application of parametric statistical methods inappropriate.

A Likert scale (Figure 2) is the attitude measurement constructed from the sum of Likert-type responses on related questions. Likert found such a summation allowed for measuring the underlying attitude that drives how a subject responds to those questions (Likert 1932). Likert scales, when properly constructed, approximate interval data and are therefore amenable to traditional parametric methods (Carifio and Perla 2008, Boone and Boone 2012).

How does a researcher know which questions are related and make up an attitude scale? In other words, how does one choose the Likert-type responses that combine to create an appropriate Likert scale? One approach is to use an existing attitudinal scale already evaluated for validity and reliability (see for example Buss and Perry 1992, Bennett and Robinson 2000). Unfortunately, these scales are not available for all research problems, although they may provide a useful starting point.

The Likert Scale: How to Build It

Exploratory Factor Analysis

For those intrepid analysts attempting to develop novel attitudinal scales, tools are required to determine which questions should be used together to form a Likert scale. Factor analysis is a powerful tool capable of modeling underlying attitudes and illuminating their existence in survey response data. Factor analysis is a multivariate data method used to understand underlying relationships or dimensions in data. These common dimensions are called factors (Dillon and Goldstein 1984, pp. 59–61). When applying factor analysis to survey data, the underlying attitudes are the factors. But in the case of a new survey without well-established Likert scales, the underlying attitudes and their relationship to the survey responses will not be known, though they should be hypothesized. With exploratory factor analysis (EFA) we can determine if there are underlying attitudes, or in this case, factors, that drive responses on survey questions. These factors can then be used to create new Likert scales.

Assume we administer a survey comprised of m questions, with each response denoted by X_i, where $i = 1, \ldots, m$. Using EFA, a realistic subset of n observable and prominent attitudes driving the survey responses is determined. Then each question response in Likert-type data is effectively a linear combination of the underlying factors ($_j$, $j = 1, \ldots, n$) weighted according to the strength of the relationship between the factor and the individual question called loading, or $_{ji}$. Together with a measurement error, this gives Equation 1:

$$X_i = \lambda_{1i}\,{}_1 + \lambda_{2i}\,{}_2 + \cdots + \lambda_{ji}\,{}_j + e_i \quad (1)$$

While the theoretical framework assumes that the underlying attitudes drive responses, these attitudes are not directly observable. We can only infer their existence. This makes sense, since how an individual feels about their military unit's culture, for example, will drive their responses to questions related to culture, and these questions will have relatively large loadings on a culture attitude scale. Similarly, feelings about culture will not drive responses to questions unrelated to culture and the corresponding loadings on nonculture-related attitude scales will be relatively small. The questions unrelated to culture will then not be included in the calculation of the Likert scale measuring culture attitudes. Therefore, the goal of EFA is to determine how many underlying attitudes exist and which attitudes

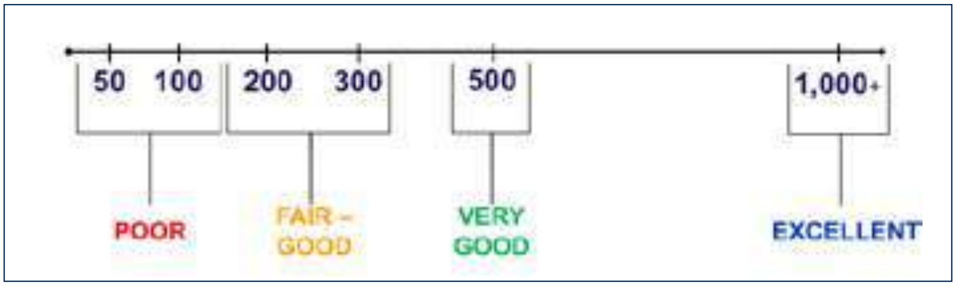

Figure 3. Sample size recommendations for exploratory factor analysis.

contribute to which set of survey questions and how much.

Fricker et al. provide background on EFA's theoretical foundations and an overview of its application to solving these problems in survey analysis, although notably without mentioning Likert (Fricker et al. 2012). The general steps to EFA outlined by Fricker et al. are:

1. Determine the number of underlying factors or attitudes (j in Equation 1)
2. Fit the j factor model in Equation 1 to estimate the factor loadings
3. Rotate the loadings (λ_{ji} in Equation 1) to simplify and find a meaningful solution

Key Considerations

While Fricker et al. describe the basic steps, there are seven considerations for building Likert scales we would like to add.

First, it would be unusual to cover statistical methods without some discussion of sample size. Figure 3 provides a summary of general sample size guidelines for EFA. Analysts should strive for greater than 200 survey respondents, although the number of questions and the nature of the data should be considered (Wilson VanVoorhis and Morgan 2007, Costello and Osborne 2005). Given that not all individuals surveyed respond, the number sampled should be adjusted based on the expected response rate.

EFA requires the creation of a correlation matrix for all questions. As noted earlier, and central to the confusion, while Likert scales can be treated as interval data, Likert-type responses are ordinal. Therefore, parametric correlations such as the common Pearson correlation are problematic because they rely on the assumption of interval data. The values for Likert-type data is restricted to the Likert-type response options. This restriction results in reduced association when measured by Pearson correlations and generates underestimated factor loadings (Holgado-Tello et al. 2010). Polychoric correlations, on the other hand, assume that the data are restricted to categories. When used for EFA of Likert-type data, polychoric correlations have been shown to provide a more accurate estimate of the true model (Equation 1) compared to Pearson correlations (Holgado-Tello et al. 2010).

Methods for fitting the EFA model to extract factor loadings include maximum likelihood (ML), principle axis factors (PAF), ordinary least squares (OLS), and weighted least squares (WLS). The ML method assumes normally distributed data and therefore some research recommends OLS when using polychoric correlations with ordinal data (Holgado-Tello et al. 2010). However, the superiority of a single method is not clear given the diverse set of recommendations in the literature (Holgado-Tello et al. 2010, Costello and Osborne 2005, Suhr 2012, Revelle 2016). We recommend using OLS and comparing the results to other extraction methods to assess model sensitivity to the method.

During extraction, loadings are rotated to simplify their structure. Rotation yields factors with predominately low loadings and questions that generally load on only one factor (Dillon and Goldstein 1984, pp. 87–89). This helps the analyst derive a meaningful interpretation of the factors without changing the general fit of the model, (e.g., the amount of variance accounted for by the items) (Costello and Osborne 2005). Of course, the method for rotating loadings must also be chosen. While Fricker et al. applied orthogonal varimax rotation to their example, we would emphasize oblique rotation such as the promax method for most survey data. Oblique rotation has the advantage of allowing for correlation between factors (i.e., attitudes), expected in the social sciences (Costello and Osborne 2005).

Once loadings are computed and rotated, a decision is required regarding which questions are associated with each factor. There is no universal prescription and this process requires a certain amount of judgment. Strong loadings are desirable, and low loadings are not, but what are low, moderate, and strong loadings? General guidelines suggest that values less than 0.4 constitute a low loading, and values greater than 0.7 are strong (Costello and Osborne 2005). In addition to strong loadings, factors extracted for analysis should include no less than three well-loaded questions (Costello and Osborne 2005, Suhr 2012).

The detected factors should be checked for reliability and goodness of fit using confirmatory factor analysis (CFA). CFA, as the name suggests, is used to confirm the existence of a hypothesized factor. There are several model metrics resulting from the CFA that may be used. Generally, an analyst should use best judgment for selection of metrics and associated cut-offs, and not simply stick to rigid rules. Some metrics and suggested cut-off values include a Cronbach's alpha measure greater than or equal to 0.6, a Tucker-Lewis index (TLI) greater than or equal to 0.95, and a root mean squared error of approximation (RMSEA) less than or equal to 0.06.[b]

Finally, the factors uncovered should have meaning to the analyst and be relevant to the research. For example, a factor discovered from a survey of military units could logically reflect an attitude about "unit leadership." It can be tempting to assume that strong loadings necessarily imply the existence of an underlying factor, but this is another case where judgment is required. The factors uncovered should have a narrative meaning of relevance to the research questions driving the study.

The Likert Scale: How to Use It

To demonstrate the use of a Likert scale, we created an attitude scale using data from the National Longitudinal Survey of American Youth (Bureau of Labor Statistics 2012).

The survey includes eight questions about the role of women, with five response options from strongly agree to strongly disagree:

1. A woman's place is in the home, not in the office or shop.
2. A wife who carries out her full family responsibilities doesn't have time for outside employment.
3. A working wife feels more useful than one who doesn't hold a job.
4. The employment of wives leads to more juvenile delinquency.
5. Employment of both parents is necessary to keep up with the high cost of living.
6. It is much better for everyone concerned if the man is the achiever outside the home and the woman takes care of the home and family.
7. Men should share the work around the house with women, such as doing dishes, cleaning, and so forth.
8. Women are much happier if they stay at home and take care of their children.

These questions were asked of subjects in 1979 and 2004. Data points with incomplete information (e.g., response refusal or "Does Not Apply") were first removed. A Likert scale to measure an attitude on women's roles was developed using EFA with the 1979 data (n = 11,873). Two potential factors were discovered underlying the eight questions. Five of the eight questions (questions 1, 2, 4, 6, and 8) were tied to one factor, with loadings between 0.58 and 0.78. This factor has enough questions and a relevant meaning (women's roles) and was therefore chosen for analysis. Several different methods for extracting the factor loadings (e.g., OLS, WLS, and ML) all resulted in the same model. Confirmatory factor analysis suggested a reliable, well-fitting Likert scale with a Cronbach's alpha of 0.78, a TLI of 0.99, and a RMSEA of 0.046. The other factor had only two well-loaded questions and was therefore not considered further.

To operationalize the Likert scale, responses for each individual on the five well-loaded questions were summed. If questions have negative loadings the responses should be reversed when calculating the score (1 becomes 5, 2 becomes 4, etc.) so that scores consistently reflect the attitude.

Using this scale, attitudes about women's roles across gender, geographic region, and time (1979 and 2004 with 5,307 respondents in both data sets) can be compared using parametric statistical methods. These results are displayed in Figure 4. The difference between genders is noticeable as is the shift toward more progressive views of women from 1979 to 2004. Hypothesis tests on the difference between years by geographic region were used to assess the change from 1979 to 2004. The increase in more progressive views of women's roles was less in the South than in the Northeast and North Central with statistical significance (α = 0.05). Also, as is noticeable in Figure 4, the move toward more progressive views from 1979 to 2004 was greater for female youths than male youths, with statistical significance (α = 0.05).

The questions that do not load onto factors (questions 3, 5, and 7) should not be discarded just because they are not suited to analysis using parametric methods. The responses may be analyzed using nonparametric methods or summarized using histograms and/or appropriate descriptive statistics (median, mode, etc.). See Figure 5 for an example.

However, analysts and decision makers should be extremely cautious when drawing conclusions from single questions, as it is possible to be misled due to question wording, misinterpretation, or other unforeseen issues.

Figure 4. Male and female youth's attitudes about women's roles in different US regions in 1979 and 2004.

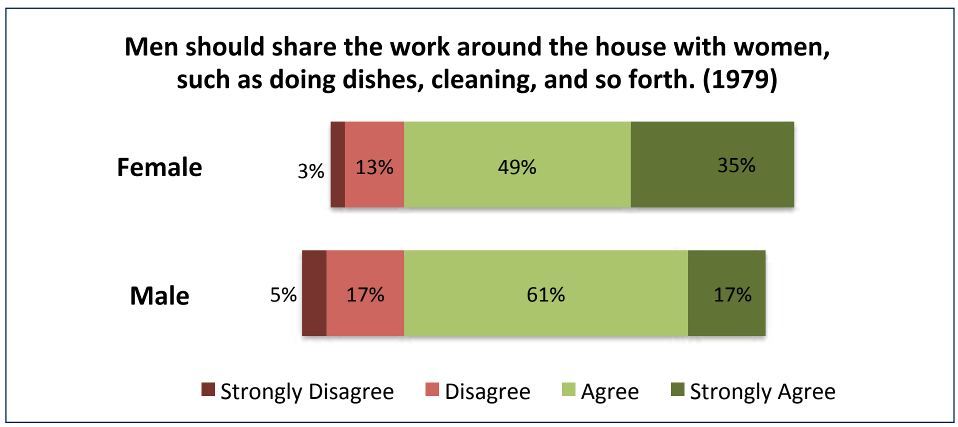

Figure 5. Example visual display of a single question.

Survey Development

Operations research analysts are sometimes prone to focus on quantitative analysis methods and spend less time on the "squishy" aspects of a study. However, proper survey development is foundational to its successful analysis. Confusing questions or an improper sampling plan could result in wasted effort. Some of the most important considerations are overviewed here. For more details, analysts may refer to resources such as the Air University Sampling and Surveying Handbook (Air University 2002) or guidance from authors such as Kulzy and Fricker (Kulzy and Fricker 2015).

The Research Question

The first step is always developing clear and well-articulated research questions. The research questions should address the purpose of the study and surveys should be used only if these research questions are answerable with a survey. The analyst should always keep the research questions in mind throughout the development and analysis of the survey.

Sampling

Sampling design should be random, considering population subgroups, such as gender, so that population inferences are possible and aligned with the research questions. Survey response rate will impact sample size and should be estimated while building the sampling plan. Additionally, weighting may be required to draw population inferences if there are response differences between subgroups. Kulzy and Fricker (2015) provide an overview of response weighting.

Question Development

Question development is critical yet extremely difficult. Best practices include avoiding excessively complex or vague questions, built-in assumptions, or questions with perceived intent. A perilous mistake is asking two questions in one, referred to as a double-barreled question. Consider this example from Likert's 1932 paper:

"Compulsory military training in all countries should be reduced but not eliminated."

A person who opposes compulsory training would disagree with the "not eliminated" portion of the statement, but a person who favors compulsory training would disagree with the "reduced" portion (Likert 1932). Because two questions are asked in one, the response cannot be used to decipher a person's attitude towards compulsory training . . . or anything else.

Once developed, questions should be thoroughly evaluated. An expert review should be used to ensure the question's meaning is appropriately conveyed. Additionally, focus groups and pilot tests may be used to assess question performance. These evaluations can identify problems early and avert issues during data collection. Finally, try to keep the survey as concise as possible.

Additional Questions

If comparisons across population subgroups are of interest, demographic questions should be included. Include enough questions to meet the needs of the research but be cautious since demographic questions can compromise anonymity.

Depending on the survey, it may also be appropriate to build screening mechanisms into the survey to detect nonsincere survey takers. Honesty questions should be used if the survey is addressing sensitive material to determine if a respondent answered honestly. Dummy questions, for example, "Choose 1, Strongly Disagree," may be used to ensure a respondent actually answered the questions instead of simply filling in responses. Any respondents answering dishonestly or failing the dummy questions should not be included in the analysis.

Those Pesky Free-form Comments: Ways to Make Sense of Them

Free-form comments are often included in surveys. These can provide a wealth of additional insight for leaders, but can also prove troublesome for analysis. To get the most out of these comments we recommend two steps: categorize and summarize. First, categorize all comments into "topics." Count how often each topic is mentioned. Then, summarize the frequencies with graphs. Finally, further summaries can be provided with text analytics. One example is a wordcloud, a visualization technique that displays the individual words sized by the frequency of use. Wordclouds can provide a nice overview of prevalent topics (Figure 6).

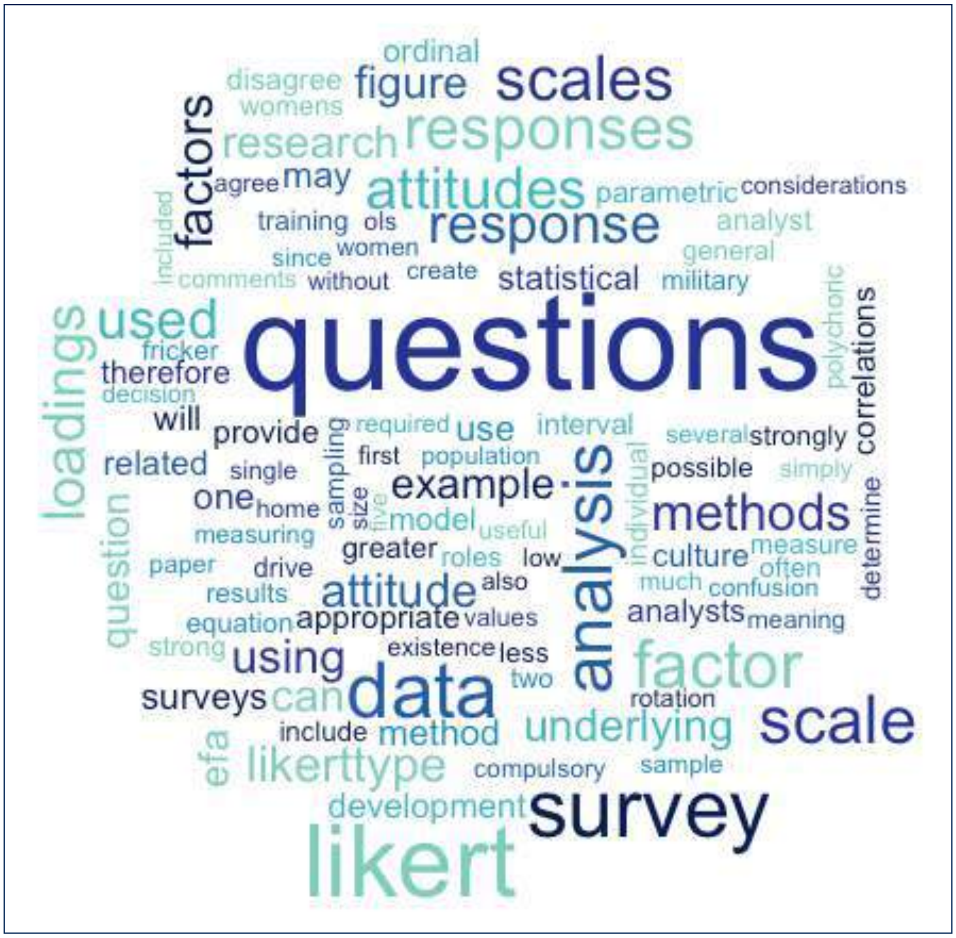

Figure 6. Example word cloud of article text.

Conclusion

Surveys provide a powerful measure of attitudes and useful insights for decision makers if they are developed and analyzed correctly. Using attitude scales, i.e., Likert scales, composed of multiple questions, provides greater insights than simply assessing responses from individual questions. To clarify the persistent confusion with Likert scales, the key is understanding the difference between Likert-type responses and Likert scales. Likert-type responses are the results of single questions and should be treated as ordinal data. Likert scales are created as the sum of multiple Likert-type responses and may be treated as interval data. Decision makers will ultimately use the information provided by surveys, so careful and precise understanding of survey development and analysis methods is essential for the operations research analyst. And, as always, it is essential that each analyst use their skill, knowledge, experience, and creativity in applying critical thinking to their unique problem.

Acknowledgements

R software (www.R-project.org) and Excel were used for the example presented in this paper. Analyses in R used packages nFactors and psych for the EFA, packages sp, maptools, ggplot2, and gridExtra for the USA maps, and packages SnowballC, tm, wordcloud, and RColorBrewer for the wordcloud. The complete R code to recreate the example is available by request from the authors.

The authors would like to thank Lt Col Eric Murphy for his thoughtful review and support.

Notes

[a] Both a 5- and 3-point response type were considered. We will not address all the various options for Likert type responses or their best use in this paper.

[b] The indices discussed here are suggested; however, other metrics may also be used for model assessment. See Hu and Bentler (1999) and Suhr (2012).

References

Air University. 2002. *Air University Sampling and Surveying Handbook: Guidelines for Planning, Organizing, and Conducting Surveys.* BiblioScholar.

Allen, I. E., and Seaman, C. A. 2007. Likert Scales and Data Analysis, *Quality Progress* (July). http://asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html.

Bennett, R., and Robinson, S. 2000. Development of a Measure of Workplace Deviance, *Journal of Applied Psychology*, Vol 85, No 3, pp. 349–360.

Boone, H., and Boone, D. 2012. Analyzing Likert Data, *Journal of Extension*, Vol 50, No 2.

Bureau of Labor Statistics. 2012. *National Longitudinal Survey of Youth 1979 Cohort, 1979–2010 (rounds 1–24)*. US Department of Labor. Produced and distributed by the Center for Human Resource Research, The Ohio State University.

Buss, A. H. and Perry, M. P. 1992. The Aggression Questionnaire, *Journal of Personality and Social Psychology*, Vol 63, pp. 452–459. doi: 10.1037/0022-3514.63.3.452.

Carifio, J., and Perla, R. 2007. Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes, *Journal of Social Sciences*, Vol 3, No 3, pp. 106–116. doi: 10.3844/jssp.2007.106.116.

Carifio, J., and Perla, R. 2008. Resolving the 50-year Debate Around Using and Misusing Likert Scales, Medical Education, Vol 42, pp. 1150–1152. doi: 10.1111/j.1365-2923.2008.03172.x.

Costello, A. B., and Osborne, J. W.. 2005. Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most from Your Analysis, *Practical Assessment Research & Evaluation*, Vol 10, No 7.

Dillon, W. R., and Goldstein M. 1984. *Multivariate Analysis: Methods and Applications*. John Wiley & Sons.

Fricker, R. D., Jr., Kulzy, W. W., and Appleget, J. A. 2012. From Data to Information: Using Factor Analysis with Survey Data, *Phalanx*, Vol 45, No 4, pp. 30–34.

Guadagnoli, E., and Velicer., W. F. 1988. Relation of Sample Size to the Stability of Component Patterns, *Psychological Bulletin*, Vol 103, No 2, pp. 265–275. doi: 10.1037/0033-2909.103.2.265.

Henry, J. R., and Smith, W. T. 2015. An Introduction to Uniformed Operations Research, *Army Sustainment*, Jan./Feb. 2015, pp. 34–35. http://www.alu.army.mil/alog/2015/JANFEB15/PDF/140106.pdf.

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., and Villa-Abad., E. 2010. Polychoric Versus Pearson Correlations in Exploratory and Confirmatory Factor Analysis of Ordinal Variable, *Quality & Quantity Journal*, Vol 44, pp. 153–166. doi: 10.1007/s11135-008-9190-y.

Hu, L., and Bentler, P. M. 1999. Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Journal Structural Equation Modeling: A Multidisciplinary Journal*, Vol 6, No 1. 1-55. doi: 10.1080/10705519909540118.

Jamieson, S. 2004. Likert Scales: How to (Ab)use Them, *Medical Education*, Vol 38, pp. 1212–1218. doi: 10.1111/j.1365-2929.2004.02012.x.

Jamieson, S. 2005. Author's reply, *Medical Education*, Vol 39, p. 971. doi: 10.1111/j.1365-2929.2005.02238.x.

Kulzy, W. W., and Fricker, R. D. Jr. 2015. The Survey Process: With an Emphasis on Survey Data Analysis, *Phalanx*, Vol 48, No 2. pp. 32–37.

Likert, R. 1932. A Technique for the Measurement of Attitudes, *Archives of Psychology*, Vol 22, No 140.

Norman, G. 2010. Likert Scales, Levels of Measurement and the "Laws" of Statistics, *Advances in Health Sciences Education*, Vol 15, pp. 625–632. doi: 10.1007/s10459-010-9222-y.

Office of the Secretary of Defense. 2015. Memorandum for the Assistant Deputy Under Secretary of the Army for Test and Evaluation (April 2). http://www.dote.osd.mil/pub/policies/2015/4-2-15_Discussion_onIncludingNeutralResponses_onSurveyQuestions(9096).pdf.

Pell, G. 2005. Uses and Misuses of Likert Scales, *Medical Education*, Vol 39, p. 970. doi: 10.1111/j.1365-2929.2005.02237.x.

Revelle, W. 2016. *Psych: Procedures for Personality and Psychological Research*, Northwestern University, Evanston, Illinois.

https://CRAN.R-project.org/package=psych Version =1.6.9.

Suhr, D. 2012. Exploratory Factor Analysis with the World Values Survey, *SAS Global Forum 2012: Statistics and Data Analysis*.

Wilson VanVoorhis, C. R., and Morgan, B. L. 2007. Understanding Power and Rules of Thumb for Determining Sample Sizes, *Tutorials in Quantitative Methods for Psychology*, Vol 3, No 2, pp. 43–50. doi: 10.20982/tqmp.03.2.p043.

About the Authors

Maj Katherine A. Batterton is an operations research analyst in the US Air Force. She has a PhD in applied mathematics with an emphasis in statistics. She has worked in operational flight test, taught math and statistics at the US Air Force Academy, and served as the chief scientist for the Air Education and Training Command Studies and Analysis Squadron. She is currently a student at the US Air Force Air Command and Staff College.

Capt Kimberly N. Hale is an operations research analyst in the US Air Force. She earned a PhD in policy analysis with a focus on quantitative techniques. She has served as a flight commander for the Air Education and Training Command Studies and Analysis Squadron and on the assessment team at the Combined Air Operations Center. She is currently the chief analyst at a fighter test squadron.