Tested Studies for Laboratory Teaching
Proceedings of the Association for Biology Laboratory Education
Volume 38, Article 53, 2017

The Care and Feeding of Survey Data

Hans D. Lemke

University of Maryland, Department of Biology, 1210 Biology-Psychology Building, College Park MD 20742 USA
(**hlemke@umd.edu**)

Surveys containing Likert scales are commonly used to assess opinions in educational research. There has been significant debate concerning the handling of data generated from these instruments. The primary source of this confusion is the difference between Likert-style items (LSI) and true Likert scales. The goals of this paper are to clarify the differences between them and provide guidelines for their analysis and presentation. Data generated from LSI are ordinal in nature and should be analyzed using appropriate nonparametric statistical tests. Likert scales require extensive testing and validation of the instrument and analysis may be performed with parametric or nonparametric statistics, depending on the structure of the data. These guidelines provide the tools for researchers to properly analyze and present their data and practitioners to evaluate and interpret published results.

Keywords: Likert scales, Likert-style items, survey, statistical analysis, ordinal data
Link to Original Poster: http://www.ableweb.org/volumes/vol-38/poster/?art=53

Surveys are a widely used tool for the collection of data concerning attitudes and opinions. The proper methods for handling these data have been debated for many years, and continue to be a source of contention (e.g., Carifio and Perla 2007). These arguments primarily focus on the appropriate methods to analyze and report the data generated from the scales. The debate arises from two primary sources: (a) a significant confusion concerning terminology, and (b) philosophical disagreements concerning the appropriateness of certain statistical tests for analyzing the data generated by these techniques. The goal of this paper is to clarify the terminology and provide recommendations for researchers who wish to use these tools.

Terminology

Likert (1932) conceived the scale bearing his name as a method for assessing attitudes. A true Likert scale is a series of items that provide a range of responses that allow the respondent to indicate agreement with a question (Clark-Carter 2010). Response choices within an item are given numeric values and are totaled to give an overall score that is related to the respondent's attitude related to a subject. The term Likert scale should be properly used for this instrument, which generates a total score based on a large number of questions. Applying the term Likert scale to single questions that use this response format creates confusion. This problem is compounded by a lack of consistent terminology in the literature for these questions, with terms such as Likert-type data, items, or scale, Likert items, Likert response format, and Likert-scale data appearing regularly. For the remainder of this paper, Likert-style items (LSI) will be used when referring to single questions. The difference between a Likert scale and an LSI is extremely important and this confusion is likely the cause of much of the debate (Carifio and Perla 2007).

This misuse of terminology is pervasive and is found in both research articles and textbooks (e.g., Sisson and Stocker 1989; Vanderstoep and Johnston 2009). Much of the debate concerning proper treatment of Likert scale generated data, such as the exchange between Jamieson and Pell in *Medical Education,* has been caused by this lack of precision in terminology (Jamieson 2004, 2005; Pell 2005). A careful assessment of both research articles and textbooks is often necessary to decide if the authors are discussing true Likert scales or LSI.

Likert-Style Items

The majority of the attention in debates on Likert-style survey data centers on the analysis of LSI, with little consensus as to what constitutes proper data analysis. As a result, researchers use a variety of statistical methodology, with the majority employing parametric techniques. In a sample of articles from the *Journal of Agricultural Education,* over half contain analysis of individual LSI from survey data (Clason and Dormody 1994). Within these, 54% reported only descriptive statistics, 24% analyzed the data further with parametric tests and 13% used non-parametric tests. Harwell's (2001)

survey of education journals revealed that 73% of research articles contained data generated from LSI and all of them were treated as interval scales. Similarly, Kaptein, Nass and Markopolous (2010) report that in the *Proceedings of the 2009 Computer Human Interaction Conference*, 46% of the papers use LSI, with 81% being analyzed with parametric and 8% with non-parametric tests. These studies demonstrate that the statistical analysis of individual items is commonplace. In their summary of the debate over the proper analysis of Likert scales, Carifio and Perla (2007) state that:

> the Likert response format is only a problem...if one analyzes each individual item separately, which one should not ever do because of the family wise error rates of repeated statistical testing...and the fact that a single item is not a scale in the sense of a measurement scale.

Though simply never analyzing data from a single item is a tempting solution to the problem, many researchers continue to use this technique. In order to devise a sensible and logical approach to the analysis of LSI, we need to identify the measurement scale that they produce before we can recommend appropriate tests.

Measurement Scale

A fundamental issue that must be addressed is whether LSI generate interval or ordinal scale data. A typical survey question asks the respondent to indicate an opinion about something using a five point scale, with the choices being labeled 'Strongly Disagree', 'Disagree', 'Neutral', 'Agree', and 'Strongly Agree'. These answers are often assigned numerical values but, based on Stevens' (1946) scales of measurement, the data are ordinal in nature. Ordinal data are categorical values represented by integers arranged in rank order, but the values lack the consistent intervals between them found between numerical values in interval data (Stevens 1946). Ordinal data also lack the arbitrary zero point necessary for interval data (Stevens 1946). For data such as survey answers, it is hard to imagine that anyone would view the spacing between categories to be equal, and Hart (1996) reports that respondents assign different weights to the differences between categories, with the largest differences at the extreme responses. The final characteristic of ordinal data is that any order preserving transformation is acceptable (Hildebrand et al. 1977). Based on these criteria, data generated from LSI should only be interpreted as ordinal in nature.

Data Analysis

Proper descriptive statistics for ordinal data include mode, median, frequencies and cross tabulations (Cohen et al. 2007). The mode or median may provide sufficient information for a quick summary of the data, but may not provide enough resolution to accurately summarize the results in a scale with only five items. The best methods for providing a complete and accurate representation of the data are frequency tables and cross tabulations (Cohen et al. 2007). For example, consider a typical teaching evaluation question where students provide a rating on a five point scale for "The instructor presented the material clearly". The data could be presented on a table containing frequencies and percentage as in Table 1. The table indicates that the majority of students were neutral on this subject but there were strong contingents who either Strongly Agreed or Strongly Disagreed.

Table 1. Example Survey Results for Question: "The Instructor Presented the Material Clearly."

Response	Frequency	Percentage
Strongly Disagree (1)	23	21.3
Disagree (2)	10	9.3
Neutral (3)	43	39.8
Agree (4)	4	3.7
Strongly Agree (5)	28	25.9
Total	108	100.0

If we assign numerical values to the categories (as is the norm) and calculate the median and mode for this data, we find that they both equal 3, or Neutral. These single values obviously do not represent the full pattern of the responses and lose information. A cross-tabulation can be easily constructed if we are further interested in how these opinions break down by gender (Table 2). This presentation clearly indicates that males have strong opinions (in both directions) about the instructor's clarity and females are Neutral regarding this subject.

As was the case with the combined data, this pattern is simple to detect from these tables, but if an attempt is made to compare the two groups using medians, the differences are undetectable. Similarly, modes provide a misleading representation of the data, with males = 5 and females = 3. This would indicate that males Strongly Agreed with the statement, though this hides the strong contingent of those who Strongly Disagreed. The best format to graphically represent the data is a 2-D bar chart (Figure 1) (Robbins 2005).

Statistics texts across disciplines recommend the use of nonparametric methods for the analysis of ordinal data (e.g., Cohen et al. 2007) because parametric methods are adequate only when data are normally

Table 2. Example Table of Cross-tabulation by Totals: Gender*The Instructor Presented the Material Clearly: Cross Tabulation

		Response					
Gender		Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)	Total
Male	Count	22	4	2	0	26	54
	% of Total	20.4	3.7	1.9	0.0	24.1	50.0
Female	Count	1	6	41	4	2	54
	% of Total	0.9	5.6	38.0	3.7	1.9	50.0
Total	Count	23	10	43	4	28	108
	% of Total	21.3	9.3	39.8	3.7	25.9	100

distributed (Zumbo and Zimmerman 1993) and the interval between data points is equal (Heeren and D'Agostino1987), conditions that rarely occur with ordinal data. A handful of researchers argue that parametric methods are appropriate (see Knapp 1990 for a summary) but their arguments are limited in scope and can lead to meaningless hypothesis testing (Marcus- Roberts and Roberts 1987). For example, Norman (2010) demonstrates that parametric methods provide acceptable results, but his conclusion is based on the analysis of only one data set. Additionally, in a comparison of t and Mann- Whitney-Wilcoxon (MWW) tests with a wide range of hypothetical distributions, deWinter and Dodou (2010) report that the two tests have similar power overall. Unfortunately their analysis ignores the fundamental requirement of both tests that the variances of the distributions of the groups are equal (Siegel and Castellan 1988).

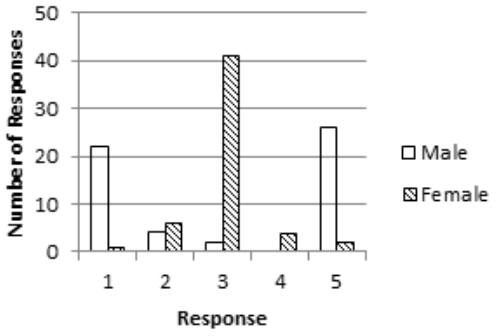

Figure 1. Student responses for Likert-style item question"The Instructor Presented the Material Clearly": Cross Tabulation" by gender. 1 = Strongly Disagree, 5 = Strongly Agree.

A myriad of tests is available for testing statistical inferences of nonparametric data. Siegel and Castellan (1988) identify these techniques as ideally suited for use in the behavioral sciences because they are free from reliance on an underlying distribution, rely on analysis of ranked data, are simple to compute, and are effective for analyzing small sample sizes. In spite of this, a bias against these tests continues to flourish among researchers and pervades the literature. Singer (1979) suggests that this may be due to fear that readers will not understand the analysis, the paper will be rejected by a journal, or the perception that these tests are inferior and only suited for simple research designs. The idea that nonparametric statistics are less powerful than their parametric counterparts is common in the literature (e.g., Armstrong 1981). The basis of this belief is unclear, but Pett (1997) suggests that this misconception results from the incorrect belief that non-normal data are inferior. A slight loss of power truly does occur when nonparametric methods are used when all of the assumptions for the parametric tests are met, but if any of the assumptions are violated, nonparametric tests have repeatedly provided substantially greater power (Cliff 1993; Conover 1999; Pett 1997). Unfortunately, this bias exists in the research community even though the conditions suited for nonparametric analyses are quite common in real world research settings.

Nonparametric tests are available for nearly any research problem. One of the most common situations encountered when analyzing LSI is the comparison of two independent samples. The MWW test is often suggested as the nonparametric alternative to the t test in these situations. As noted previously, one of the assumptions of this test is that the samples come from groups with the equal variance (Pett 1997). In cases of unequal variance, the Robust Rank-Order (RRO) or Kolmogorov-Smirnov (KS) tests may be employed (Siegel and Castellan 1988). Caution should be used when employing the RRO test as it only takes into account unequal variance, while the KS test is more flexible as it accounts for all differences in the distributions.

Another significant factor to consider when analyzing LSI is that when multiple, related tests are

conducted; the experimentwise error rate drastically increases (Sokal and Rohlf 1995). For example, consider a 10 question survey where 10 t tests are conducted on related data, with each test evaluated separately at $\alpha = .05$. This procedure is likely to produce an incorrect interpretation of the results. At this level of significance, the likelihood of committing a Type I error, or falsely rejecting the null hypothesis, in one comparison is 5%. However, when conducting multiple, related tests, the experimentwise error rate needs to be calculated (Sokal and Rohlf 1995); resulting in $\alpha = .4$ in this case. This means that at the same critical value, there is now a 40% chance of making a Type I error in the experiment. Analyzing 20 questions this way increases the experimentwise error rate to 64%. To avoid this, the use of nonparametric analysis analogous to ANOVA, such as the Kruskal-Wallis test, is recommended (Sokal and Rohlf 1995).

Likert Scales

Information concerning true Likert scales is relatively hard to come by in statistics textbooks. Though analysis of the data obtained using these tools is relatively straight forward, the construction of a proper scale is rather laborious.

Construction

Clark-Carter (2010) and Blaikie (2003) provide extensive guidelines for the construction and validation of a Likert scale. Clark-Carter (2010) suggests using at least 20 questions with either a 5- or 7-point scale. Further, about half of the statements should be worded in the opposite direction from the rest, allowing for internal validation of the scale. Attitude measuring devices, such as Likert scales, require validation to determine how many dimensions the scale is measuring and if each item is receiving an appropriate range of answers (Clark-Carter 2010). Blaikie (2003) recommends that the scale undergo a period of pre-testing and be subjected to a 2-step item analysis before it is employed in a research setting. The first is an item-to-item correlation, which allows the researcher to determine if the scale is actually measuring one or multiple dimensions. The second step, an item-to- total correlation, checks the discriminatory power of each dimension, so that the researcher can be sure that differences between them are indicative of the score on the total scale. Following this verification process, the scale's reliability should be determined. Blaikie (2003) recommends using Cronbach's α, but Gaderman, Guhn, and Zumbo (2012) caution that this test is highly affected by outliers and is not suited for use with ordinal data. They recommend the use of the ordinal coefficient α instead. Finally, a factor analysis should be conducted to determine if any items should be excluded. This process should be undertaken before the use of the survey and noted in the Methods section of a research report.

Unfortunately, the validation of instruments used to measure performance and attitudes is frequently lacking in research methodology (Phipps and Merisotis 1999).

Measurement Scale

The scores generated from a properly constructed Likert scale can be assumed to be on an interval scale (Blaikie 2003). The literature contains few arguments to the contrary; however those authors who do claim otherwise are apparently referring to LSI when they make their cases (e.g., Jamieson 2004).

Data Analysis

Data derived from Likert scales can be analyzed using parametric techniques, provided that the assumptions for the tests are met. The primary considerations are random sampling, independence, normality, and homogeneity of variance between samples (Sokal and Rohlf 1995). These assumptions need to be tested using methods such as Kolmogrov-Smirnov or Shapiro-Wilk tests (Field, 2009) and the results reported. How closely the data need to conform is a matter of debate. Carifio and Perla (2007) argue that F tests are robust against deviations from the assumption of normality. They cite numerous papers that use Monte Carlo studies on simulated data that demonstrate this robustness. However, in making their arguments, they leave out a long history of studies focusing on the problems associated with assuming normality when it is not present. Micceri (1989) provides an excellent summary of this research, dating back to the 1890s. In addition, real world data often behave quite differently than do artificially generated data, and the simulations do not examine lumpiness or multimodality, which are common in practice (Micceri 1989). Nanna and Sawilowsky (1998) examined a large data set generated from an 18 item, 7-point Likert scale to test the relative power of Wilcoxon rank-sum and t tests. The Wilcoxon rank-sum test was more powerful for nearly all sample sizes and alpha levels. This was true for tests using both individual items from the scale (LSI) and the overall score. Micceri (1989) examined over 400 score distributions from a range of psychometric and achievement/ability tests to determine how often real-world data conform to guidelines for normality. He found that none of the data sets passed all of the tests for normality and that few even came close. He recommends that, due to its rarity, testing for normality is a waste of time as it would effectively only occur by chance. Following his logic, researchers should simply use nonparametric tests as a rule. This would most likely be unsatisfying to many researchers so, at a minimum, these findings should caution researchers from assuming normality in data.

The choice of descriptive statistics depends on the nature of the data. If the data conform to a normal distribution then mean and standard deviation are appropriate. However, the mean is heavily affected by outliers and changes in the shape of the distribution so, in this case, the median may be more appropriate (Sokal and Rohlf 1995). If the median is chosen, then interquartile ranges should be used to indicate the dispersion of the data (Sokal and Rohlf 1995). A boxplot can be used to produce a visual representation of the median and interquartile ranges (Field 2009).

The choices for a statistical test to use when analyzing this Likert scale data are governed by the researcher's decision as to whether or not the distribution is parametric. If parametric tests are used, it is advisable to list the assumptions made in making that decision.

Discussion

In concluding his discussion on the debate concerning the analysis of Likert scales, Norman (2010) stated that "the controversy can cease (but likely won't)." The arguments presented here are also unlikely to bring this controversy to a close. Hopefully this review will help to clarify the main issues surrounding the design, analysis and reporting of data generated through surveys. The complications involved in the construction and verification of true Likert scales will likely lead to the continued use of LSI. The methods presented here allow researchers to rigorously analyze these data and avoid common pitfalls found in the literature. This requires the use of techniques, such as the use of nonparametric statistics and cross tabulations that are outside the comfort zone of many people, but fortunately are relatively simple to learn. Norman (2010) makes the argument that if "we have to prove that our data are exactly normally distributed, then we can effectively trash about 75% of our research on educational, health status and quality of life assessment." This is probably hyperbole, but his statement does serve as a caution for those who are evaluating previously published work. Readers should look carefully at the data before using it. Data can be re- analyzed using appropriate techniques if they are well reported. At a minimum, appropriate α levels can be calculated for instances involving multiple comparisons.

Though these arguments are unlikely to sway the minds of those who believe that ordinal data can be dealt with as if it were interval in nature, this review should serve as a notice to those conducting research or reading reports that if parametric analyses are used, it should be with caution.

Cited References

Armstrong GD. 1981. Parametric statistics and ordinal data: A pervasive misconception. Nurs Res. 30:60-62.

Blaikie N. 2003. Analyzing quantitative data: From description to explanation. London (England): Sage.

Carifio J, Perla RJ. 2007. Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. J Soc Sci. 3:106-116.

Clark-Carter D. 2010. Quantitative psychological research: The complete student's companion. Hove (England): Psychology Press.

Clason DL, Dormody TJ. 1994. Analyzing data measured by individual Likert-type items. J Agric Educ. 35:31-35.

Cliff N., 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. Psychol Bull. 114:494-509.

Cohen L, Manion L, Morrison K. 2007. Research methods in education. 6th ed. London (England): Routledge.

Conover WJ. 1999. Practical nonparametric statistics. 3rd ed. New York (NY): John Wiley and Sons.

de Winter JCF, Dodou D. 2010. Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. Prac Assess Res Eval. 15(11):1-12.

Field, A. 2009. Discovering statistics using SPSS. 3rd ed. Los Angeles (CA): Sage.

Gaderman AM, Guhn M, Zumbo BD. 2012. Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. Prac Assess Res Eval. 17 (3):1-13.

Hart MC. 1996. Improving the dissemination of SERVQUAL by using magnitude scaling. In: Kanji GK, editor. Total quality management in action. London (England): Chapman and Hall. p. 267-270.

Harwell M. 2001. Future directions and suggestions for improving the use of statistical methods in educational research. Paper presented at the Annual Meeting of the American Educational Research Association; Seattle (WA). Retrieved from ERIC database.

Heeren T, D'Agostino R. 1987. Robustness of the two independent samples t-test when applied to ordinal scale data. Stats in Med. 6:79-90.

Hildebrand DK, Laing JD, Rosenthal H. 1977. Analysis of ordinal data. Beverly Hills (CA): Sage.

Jamieson S. 2004. Likert scales: How to (ab)use them. Med Educ. 38:1217-1218.

Jamieson S. 2005. Author's reply. Med Educ. 39:971.

Kaptein M, Nass C, Markopolous P. 2010. Powerful and consistent analysis of Likert-type rating scales. Paper presented at 28th International Conference on Human Factors in Computing Systems: Atlanta (GA). Retrieved from http://portal.acm.org/citation.cfm?doid=1753326.1753686

Knapp TR. 1990. Treating ordinal scales as interval scales: An attempt to resolve the controversy. Nurs Res. 39:121-123.

Likert R. 1932. A technique for the measurement of attitudes. Arch Psychol. 140:1-55.

Marcus-Roberts HM, Roberts FS. 1987. Meaningless statistics. J Educ Stat. 12: 383-394.

Micceri, T. 1989. The unicorn, the normal curve, and other improbable creatures. Psychol Bull. 105:156-166.

Nanna MJ, Sawilosky SS. 1998. Analysis of Likert scale data in disability and medical rehabilitation research. Psychol Meth. 3: 55-67.

Norman G. 2010. Likert scales, levels of measurement and the "laws" of statistics. Adv Health Sci Educ. 15: 625-632.

Pell G. 2005. Use and misuse of Likert scales. Med Educ. 39: 970.

Pett MA. 1997. Nonparametric statistics for health care research: Statistics for small samples and unusual distributions. Thousand Oaks (CA): Sage.

Phipps R, Merisotis J. 1999. What's the difference? A review of contemporary research on the effectiveness of distance learning in higher education. Retrieved from EBSCOhost.

Robbins NB. 2005. Creating more effective graphs. Hoboken (NJ): John Wiley and Sons.

Siegel S, Castellan NJ Jr. 1988. Nonparametric statistics for the behavioral sciences. Boston (MA): McGraw Hill.

Singer B. 1979. Distribution-free methods for non-parametric problems: A classified and selected bibliography. Brit J Math Stat Psychol. 32:1-60.

Sisson, DV, Stocker HR. 1989. Research corner: Analyzing and interpreting Likert-type survey data. Delta Pi Epsilon J. 31: 81-85.

Sokal RR, Rohlf FJ. 1995. Biometry: The principles and practice of statistics in biological research. New York (NY): W.H. Freeman and Company.

Stevens SS. 1946. On the theory of the scales of measurement. Science 103: 677-680.

Vanderstoep SW, Johnston DD. 2009. Research methods for everyday life: Blending qualitative and quantitative approaches. San Francisco (CA): Jossey-Bass.

Zumbo BD, Zimmerman DW. 1993. Is the selection of statistical methods governed by level of measurement? Can Psychol. 34: 390-400.

Acknowledgments

Thank you to the reviewers who have provided helpful feedback on this project throughout the years.

About the Author

Hans is the Lab Coordinator for the Principles of Ecology and Evolution Lab (BSCI161) at the University of Maryland. He holds a B.A. in Biology from St. Mary's College of Maryland, an M.S. in Entomology from the University of Maryland, and an M.D.E. in Distance Education from University of Maryland, University College. In addition to coordinating the labs, he helps to teach courses on experimental design. Current research focuses on educational outcomes in laboratory and online settings and a survey of tiny Miocene shark and ray fossils found along the Chesapeake Bay.

Mission, Review Process and Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit **http://www.ableweb.org/.**

Papers published in *Tested Studies for Laboratory Teaching: Peer-Reviewed Proceedings of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Citing This Article

Lemke HD. 2017. The Care and Feeding of Survey Data. Article 53 In: McMahon K, editor. Tested studies for laboratory teaching. Volume 38. Proceedings of the 38th Conference of the Association for Biology Laboratory Education (ABLE). **http://www.ableweb.org/volumes/vol-38/?art=53**