Journal of Agricultural Education
Volume 35, Number 4, pp. 31-35
DOI: 10.5032/jae.1994.04031

Analyzing Data Measured by Individual Likert-Type Items

Dennis L. Clason, Associate Professor
Thomas J. Dormody, Assistant Professor
New Mexico State University

Using individual (not summated) Likert-type items (questions) as measurement tools is common in agricultural education research. The Journal of Agricultural Education published 188 research articles in Volumes 27 through 32. Responses to individual Likert-type items on measurement instruments were analyzed in 95, or more than half, of these articles. After reviewing the articles analyzing individual Likert-type items, 51 (54%) reported only descriptive statistics (e.g., means, standard deviations, frequencies/percentages by category). Paired Likert-type items or sets of items were compared using nonparametric statistical techniques (e.g., chi-square homogeneity tests, Mann-Whitney-Wilcoxon U tests, Kruskal-Wallis analysis of variance tests) in 12 (13%) of the articles. Means for paired Likert-type items were compared using parametric statistical procedures (e.g. t-tests or analysis of variance F-tests) in 32 (34%) of the articles.

These data are introduced to illustrate that a variety of statistical methods are being used to analyze data from individual Likert-type items in the Journal of Agricultural Education. Which are most appropriate? On what bases should we make our choice of methods? This paper explores these questions.

Likert Scales and Likert-Type Items

Likert (1932) proposed a summated scale for the assessment of survey respondent's attitudes. Individual items in Likert's sample scale had five response alternatives: Strongly approve, Approve, Undecided, Disapprove, and Strongly disapprove. Likert noted that descriptors could be anything -- it is not necessary to have negative and positive responses. He implies that the number of alternatives is also open to manipulation. Indeed, we see contemporary work using many classifications besides the traditional five point classifications; some researchers use an even number of categories, deleting the neutral response. The 95 articles analyzing individual Likert-type items in Volumes 27 through 32 of the Journal of Agricultural Education used response alternatives ranging from three to eight (or more) points (Table 1).

Table 1. Journal of Agricultural Education Volumes 27 through 32: Number of Response Alternatives Used in Likert-Type Items

Number of Response Alternatives	Number of Studies
3	7
4	28
5	67
6	12
7	7
8 to 99	11

Likert's original work assumed an attitude scale would first be pilot tested for reliability assessment of the individual items. This reliability assessment might use the correlation between the item score and the total or use a split-half procedure. In any event, the items not correlated with the total would be discarded. Subsequent data would be summarized using the totals. He apparently did not consider the possibility that individual items might be analyzed. Indeed, there is some confusion over whether a Likert scale refers to a summation of the item scores or if it refers to the number of response alternatives in individual items. Likert's monograph (1932) makes it clear he never intended for the five-point, seven-point or other response alternatives to be the scale. Although they represent a scale of sorts when coded, they are not a Likert (summated) scale. To distinguish individual items from the summated Likert scale, we will refer to the individual items as Likert-type items.

Likert scaling presumes the existence of an underlying (or latent or natural) continuous variable whose value characterizes the respondents' attitudes and opinions. If it were possible to measure the latent variable directly, the measurement scale would be, at best, an interval scale. Goldstein and Hersen (1984) state this clearly:

> The level of scaling obtained from the Likert procedure is rather difficult to determine. The scale is clearly at least ordinal. Those persons with the higher level properties in the natural variable are expected to get higher scores than those persons from lower properties. . . In order to achieve an interval scale, the properties on the scale variable have to correspond to differences in the trait on the natural variable. Since it seems unlikely that the categories formed by the misalignment of the five responses will all be equal, the interval scale assumption seems unlikely. (p. 52)

It is probable the Likert scale will be ordinal, but in any event, the population could be totally ordered by the magnitude of the latent variable. A single Likert-type item asks the respondent to which of several ordered alternatives they belong. Each Likert-type item provides a discrete approximation of the continuous latent variable. A proper analysis of single items from Likert scales should acknowledge the discrete nature of the response. We should probably note here that we are not addressing the issue of parametric versus nonparametric analysis of the Likert scale scores, as addressed by Dawis (1987) and Adams, Fagot and Robinson (1965). We are examining the proper analysis of single Likert-type items only.

Ignoring the discrete nature of the response can lead to inferential errors. An extreme case will serve as an illustration. Suppose a particular item on an attitude questionnaire is "Student evaluations are an excellent indicator of a teacher's classroom performance." Let the responses be Strongly agree (5) to Strongly disagree (1). Let us further suppose we want to compare the responses of two populations to this item, say teachers of agriculture and state supervisors. We observe means of 3 -- "Undecided" for both populations. According to a t-test, *or any test focused on location*, these two populations are identical. These tests ignore the discrete nature of the response, and use the item response to approximate the unobservable latent variable. This analysis assumes the existence of an invertible function that uniquely maps the latent variable into the Likert-type item. It's sad, but no such function can exist, for such a function can take on only five (or seven or nine . . .) unique values, while there are infinitely many values of the latent variable. Moreover, the existence of such a function implies a strongly homogeneous interpretation of the statement in the population.

If the analyst acknowledges the discrete nature of the observations, the data will be summarized as counts (or percentages) occurring in the various response categories. To return to the previous example, the analyst might have found the following observed proportions:

	Response				
	1	2	3	4	5
Teachers	0.20	0.20	0.20	0.20	0.20
Supervisors	0.50	0.00	0.00	0.00	0.50

Both groups have mean scores of 3.00, or neutral responses, but only the most naive analyst would claim these populations are similar.

The Multinomial Distribution

A generalization of the binomial (how many successes in n independent identical trials) provides the theoretical framework needed to analyze Likert-type items. The generalization involves extending the counting from two bins (say, success and failure) to c bins. Every observation can be correctly placed into exactly one of the c bins. The usual five-point Likert-type item can be represented as a c =5 multinomial. We assume the probability that a randomly selected individual falls into bin "j" is π_j. The probability function for a single observation is:

$$m(\boldsymbol{x}|1,\boldsymbol{\pi}) = \pi_1^{x_1}\pi_2^{x_2}\cdots\pi_c^{x_c} = \prod_{j=1}^{c}\pi_j^{x_j}, \quad \text{for } x_j\in\{0,1\}.$$

Following a development similar to that used to obtain the binomial distribution from the Bernoulli distribution (Mood, Graybill, & Boes, 1974), we find the probability function for sample counts is:

$$m(\boldsymbol{x}|n,\boldsymbol{\pi}) = \binom{n}{x_1,x_2,\ldots,x_c}\prod_{j=1}^{c}\pi_j^{x_j},$$

$$\text{for } x_j\in\{0,\ 1,\ 2,\ \ldots,\ n\}$$

$$\text{and } \Sigma x_j = n.$$

Just as the best estimate of P[Success] in a binomial experiment is π= X/n, the best estimator of

the multinomial population proportion in category "j" is $\pi_j = X_j/n$. Statistical procedures that make full use of the discrete nature of a Likert-type item will be based on either the cell counts or the cell proportions, because these are minimally sufficient statistics for the problem. Confidence intervals (and tests) that are not based on sufficient statistics necessarily discard information (Mood, et al., 1974).

Count-Based Analyses

Generally, the question of interest is of the form "Does subpopulation 1 differ from subpopulation 2?" Because the sample sizes are fixed by design (assuming a stratified random sample), the null hypothesis stated in terms of multinomial parameters is $H_0 : \pi_1 = \pi_2$, the alternative being inequality. There are two appropriate test statistics that can be used here (Fienberg, 1977). The best known statistic is Pearson's X^2; the other statistic is based on the generalized likelihood ratio. Pearson's X^2 is calculated using the familiar formula:

$$X^2 = \sum_{i=1}^{r} \sum_{i=1}^{c} \frac{(X_{ij} - \frac{X_{i+}X_{+j}}{X_{++}})^2}{\frac{X_{i+}X_{+j}}{X_{++}}}$$

Often written as

$$X^2 = \Sigma \frac{(Observed-Expected)^2}{Expected}.$$

Here X_{i+}, X_{+j}, and X_{++} respectively indicate row (population), and column (response category), and grand totals of the cell counts. The likelihood ratio test statistic is calculated using the formula:

$$G^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} X_{ij} \log_e \frac{X_{ij}}{\frac{X_{i+}X_{+j}}{X_{++}}},$$

or:

$$G^2 = 2 \Sigma\, Observed \times \log_e \frac{Observed}{Expected}$$

If the multinomial model is appropriate and the sample sizes are large, these statistics are usually in close agreement. The derivation of the sampling distributions in each case relies on asymptotic (large sample) arguments, so neither should be relied on to be accurate for small samples in the absence of empirical research. Work in this area (Chapman, 1976) suggests the average approximate significance level (ASL) for G^2 may be closer to the true (multinomial) significance level than the ASL for X^2. In other words, ASLs for X^2 are less biased and ASLs for G^2 are more precise. This creates a quandary for the researcher who needs a test statistic for a problem with a small sample. It appears at this time that the choice is a matter of personal preference.

Measures of Association

Sometimes the question is one of association, e.g., "Does subpopulation 1 tend to have higher (or lower) scoring for this item than subpopulation 2?" Answering this question calls for a measure of association rather than a formal test for differences. Just as in the testing case, the discrete and ordinal nature of the response restricts the measures that can be used. Pearson's **r** is particularly inappropriate because it is influenced by the range used in response coding. That is, two investigators could take the same data and, by choosing to code 1 to 25 rather than 1 to 5, obtain completely different results for **r**. Likert (1932) makes it clear that origin and width of scale are not relevant. Pearson's product moment correlation is, therefore, useless because it is sensitive to the choice of scale. There are scoring modifications that correct this problem and make **r** usable in situations where both variables are ordinal.

Kendall's tau coefficient is an appropriate choice when the subpopulations are ordinal in some sense, say as age-group breakouts (Goodman and Kruskal, 1979). Other appropriate choices for ordinal classifications include the Spearman rank correlation and variants on the tau coefficient (e.g., gamma coefficient; Somer's D coefficient). The choices are more limited when the classification variable is not ordered. One can choose from the family of X^2 based measures, e.g., the Phi coefficient, the contingency coefficient and Cramer's V, or the lambda coefficient or the

uncertainty coefficient (U). Lambda and U are both predictive measures, quantifying the improvement in predicting one variable on the basis of knowledge of the second. Because of their definition they are probably not good general choices, but can be very useful when appropriate. The three X^2 measures are modifications to bring the X^2 statistic into the range [0, 1].

Other Considerations

Some statisticians have no problem with analyzing individual Likert-type items using t-tests or other parametric procedures (Sisson & Stocker, 1989), provided the primary interest is in location only. If the survey process produces order and normality, normal theory procedures can be employed regardless of the attained measurement level. We do not dispute the logic, but disagree with the premise when discussing Likert-type items. It is difficult to see how normally distributed data can arise in a single Likert-type item. The data will frequently be skewed, and often these items do not capture the true limits of the attitude. An individual item will often produce distributions showing a floor or ceiling effect -- respondents choosing the lowest or highest available alternative. In these situations, the true mean for a Likert-type item may not be measurable because of limitations imposed. Tests of means in these situations are problematic: mean differences become more a function of sample size than respondent attitude. Summative scales provide one path out of this quagmire.

Checking for normality before selecting an inference procedure necessitates selecting the inference procedure post hoc. Many statisticians argue against post hoc selection, they feel researchers should determine appropriate inferences procedures in the planning stage. Adherents of this viewpoint are concerned with inferential validity when the data analyst shops for a statistic showing differences. Another point favoring this view is the low power of available normality tests when sample sizes are small, and the excessive power of these tests for large samples. As a model, the normal distribution probably describes no population completely -- it is no surprise that the data are never exactly normal. There are no hard and fast rules for deciding how normal is normal enough, so researchers operating under the order and normality assumptions will necessarily make their decisions with different criteria.

Many of the 95 studies reported in Volumes 27 through 32 of the Journal of Agricultural Education using Likert-type items as measurement tools analyzed many subsets (pairs, triplets) of items. Even if all the items in some sets met the criteria for normality, it is confusing to analyze some subsets with parametric techniques while analyzing qualitatively similar subsets using a X^2 approach. To be consistent in answering research questions, the analysis ought to be the same for all similar subsets.

Summary

It is not a question of right and wrong ways to analyze data from Likert-type items. The question is more directed to answering the research questions meaningfully. Some techniques answer meaningful questions completely, others ignore aspects of the problem. Adams, Fagot, and Robinson (1965) said:

> Nothing is wrong per se in applying any statistical operation to measurements of given scale, but what may be wrong, depending on what is said about the results of these applications, is that the statement about them will not be empirically meaningful or else that it is not scientifically significant. (p. 100)

If we want to know if two or more populations are different or if the same population is different on several measures and Likert-type items are generating the data, methods focusing on location parameters may oversimplify the analysis. Statistical procedures that meaningfully answer the research questions, maintain the richness of the data, and are not subject to scaling debates should be the methods of choice in analyzing Likert-type items.

References

Adams, Fagot, & Robinson. (1965). A theory of appropriate statistics. Psychometrika, 30(2): 99-127.

Chapman, J.W. (1976). A comparison of the X^2, -2 log R, and multinomial probability criteria for significance tests when expected frequencies are small. Journal of American Statistical Association, 71: 854-863.

Dawis, R.V. (1987). Scale construction. Journal of Counseling Psychology, 34: 481-489.
Fienberg, S.E. (1977). The Analysis of Cross-classified Categorical Data. Cambridge, MA: MIT Press.
Goodman, LA., & Kruskal, W.H. (1979). Measures of Association for Cross-classification. New York: Springer-Verlag.
Goldstein, G., & Hersen, M. (1984). Handbook of Psychological Assessment. New York: Pergamon Press.
Likert, R. (1932). A Technique for the Measurement of Attitudes. New York: Archives of Psychology.
Mood, A.M., Graybill, F.A., Boes, D.C. (1971). An Introduction to the Theory of Statistics. New York: McGraw-Hill.
Sisson, D.A., & Stocker, H.R. (1989). Analyzing and interpreting Likert-type survey data. The Delta Pi Epsilon Journal, 31(2): 81-85.