Journal of Educational Technology Development and Exchange (JETDE)

Volume 13 | Issue 2

3-25-2021

Methods to Analyze Likert-Type Data in Educational Technology Research

Li-Ting Chen
University of Nevada, Reno, litingc@unr.edu

Leping Liu
University of Nevada, Reno, liu@unr.edu

Recommended Citation

Chen, Li-Ting and Liu, Leping (2021) "Methods to Analyze Likert-Type Data in Educational Technology Research," *Journal of Educational Technology Development and Exchange (JETDE)*: Vol. 13 : Iss. 2 , Article 3.
DOI: 10.18785/jetde.1302.04
Available at: https://aquila.usm.edu/jetde/vol13/iss2/3

Chen, L.T., & Liu, L.(2020). Methods to Analyze Likert-Type Data in Educational Technology Research. *Journal of Educational Technology Development and Exchange*, ***13***(2), 39-60

Methods to Analyze Likert-Type Data in Educational Technology Research

Li-Ting Chen
University of Nevada, Reno
Leping Liu
University of Nevada, Reno

Abstract*: Likert-type items are commonly used in education and related fields to measure attitudes and opinions. Yet there is no consensus on how to analyze data collected from these items. In this paper, we first provided a synthesis of the existing literature on methods to analyze Likert-type data and computing tools for these methods. Secondly, to examine the use and analysis of Likert-type data in the field of educational technology, we reviewed 424 articles that were published in the journal Educational Technology Research and Development between 2016 and 2020. Our review showed that about 50% of the articles reported Likert-type data. A total of 139 articles used Likert-type data as a dependent variable, among which 86% employed parametric methods to analyze the data. In addition, less than 3% of the 139 articles used an ordered probit/logit model, transformation, or strategy for rescaling Likert-type data to interval data to perform statistical analysis. Finally, to empower educational technology researchers to handle Likert-type data effectively, we concluded the paper with our suggestions and insight regarding alternative strategies and methods.*

Keywords: Likert, ordinal data, research method, self-report measures, rescaling, nonparametric, ordered probit, ordered logit, robust

1. Introduction

Questionnaires that ask individuals (e.g., students, parents, or teachers) to rate their attitudes and opinions about statements using Likert-type items (Likert, 1932) are common in education and related fields (Antonialli et al., 2017; Carifio & Perla, 2007; Edmondson, 2005; Harwell & Gatti, 2001; Liddell & Kruschke, 2018; Potvin & Hasni, 2014; Tsui, 1997). For example, a teacher may be asked to respond to the statement "I can learn technology easily" using the five response options: 0 = strongly disagree, 1 = disagree, 2 = neither agree nor disagree, 3 = agree, 4 = strongly agree. Although the options are labelled using numbers, the numerals only indicate orders. They do not necessarily imply that distances between two adjacent options are equal. That is, the distances between 0 and 1, 1 and 2, 2 and 3, and 3 and 4 may be different.

According to Stevens (1946), there are four types of measurements (nominal, ordinal, interval, and ratio) and the types of measurements are determined by their basic empirical operations. Nominal measurement consists of category labels (e.g., numbers or symbols) that can be assigned to observations (or individuals) so that those with different labels are not equivalent. With an ordinal measurement, category labels are assigned to observations to rank and order them with respect to one another. Using an interval measurement, numbers are assigned to observations. The numbers have the property of order, and equal differences between any two adjacent numbers reflect equal magnitude. All the properties of an interval measurement apply to a ratio measurement, and in addition, there is a true zero point for a ratio measurement to reflect the absence of the measured characteristic. Stevens emphasized the permissible and impermissible transformations for numbers yielded from the four types of measurements. Permissible transformations are transformations that maintain the same meanings of the numerals assigned to observations. Permissible transformations for ordinal data, such as Likert-type data, are monotonic transformations, or positive linear transformations, but not one-to-one substitutions. Permissible transformations for interval data are positive linear transformations but not monotonic transformations or one-to-one substitutions. Following the prescription, Stevens (1946, 1955, 1968) urged researchers to attend to the type of measurement. Stevens stated

> The *ordinal scale* arises from the operation of rank-ordering. … most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales, for these statistics imply a knowledge of something more than the relative rank-order of data (Stevens, 1946, p. 679).

In fact, contrary to Stevens's suggestion, analyzing Likert-type data using statistical methods that require interval or ratio data is a common practice. These statistical methods include but are not limited to the *t* test, *F* test, Pearson's correlation, and ordinary least squares regression. These methods are also called parametric methods. In a commentary article that published in *Medical Education*, Carifio and Perla (2008) pointed out "How Likert type measurement scales should be appropriately used and analysed has been debated for over 50 years" (p. 1150). Just what are the discussions related to the use and analysis of Likert-type data? What analysis strategies have been proposed in the literature for dealing with Likert-type data? How prevalent is Likert-type data used in studies

in the field of education technology? What are the analysis strategies used by educational technology researchers to handle Likert-type data? In this paper, we address these questions. Specifically, there are three aims of this paper:

1. to summarize the strategies and methods to use and analyze Likert-type data from the literature;
2. to investigate the use and analysis of Likert-type data in educational technology research; and
3. to provide suggestions for educational technology researchers to handle Likert-type data.

2. Literature Review: Strategies and Methods to Analyze Likert-Type Data

Likert summative attitude scales were first introduced by Rensis Likert in the 1930s. Likert (1932) suggested that response options to several statements on Likert summative attitude scales can be written as 1 = strongly disapprove to 5 = strongly approve. Such response options are widely used to measure attitudes and opinions. In this paper, we used the term *Likert-type data* to refer to data collected from such response format, and the term *Likert-type item* to refer to a statement with such response format to measure individuals' attitudes and opinions. In the field of educational technology, researchers may ask online students to respond to statements on the 34-item Community of Inquiry framework survey with five response options to understand students' perception of cognitive presence, social presence, and teaching presence (Arbaugh et al., 2008). To understand pre-service teachers' self-assessment of Technological Pedagogical Content Knowledge (TPACK) and related knowledge domains included in the TPACK framework, researchers may also ask pre-service teachers to respond to statements on the 47-item Pre-service Teacher Technological Pedagogical Content Knowledge Instrument with five response options (Schmidt et al., 2009).

One potential problem with using parametric methods for Likert-type data is about the normality assumption, which requires continuous/interval data. Scores yielded from Likert-type items are discrete and ordinal in nature. For example, when five items are used to measure students' perceived competence in information and communication technology and all five items are rated using four response options (Areepattamannil & Santos, 2019), ranging from 1 (strongly disagree) to 4 (strongly agree), there are 16 possible total scores with the minimum score of 5 and maximum score of 20. Furthermore, when computing the total scores with weighting each item equally, it ignores the unique characteristics of each item (Harwell & Gatti, 2001).

While some researchers think it is crucial to consider the type of measurement when using a statistical method (Jamieson, 2004; Kuzon et al., 1996; Siegel, 1956; Sprinthall, 2012), others care less about the type of measurement (Carifio & Perla, 2007; Howell, 2013; Lord, 1953; Norman, 2010; Velleman & Wilkinson, 1993; Zimmerman, 2011). Kuzon et al. (1996) referred using parametric methods for ordinal data as the first sin of the seven deadly sins of statistical analysis and suggested "*to avoid committing Sin 1, for nominal or ordinal scaled data, use nonparametric statistical analysis*" (p. 266). In contrast, Carifio and Perla (2007) presented the top ten myths about "Likert scales" and wrote "Myth 6—Because Likert scales are ordinal-level scales, only non-parametric statistical tests should be used with them" (p. 114). Parametric methods differ from nonparametric methods in making assumptions

about the parameters of the population distribution from which the sample is drawn, such as normality assumption, for valid inferences. When assumptions are not met, estimates of parameters may be overestimated or underestimated, statistical tests may be more likely to reject a true null hypothesis (Type I error) than a nominal alpha (e.g., .05), or fail to reject a false null hypothesis (Type II error). The most common nonparametric methods are the Mann-Whitney *U* test (a nonparametric counterpart of independent-samples *t* test), Wilcoxon's rank sum test (a nonparametric counterpart of independent-samples *t* test), Wilcoxon's matched-paired signed-rank test (a nonparametric counterpart of dependent-samples *t* test), Kruskal-Wallis test (a nonparametric counterpart of *F* test), Spearman rank correlation coefficient (a nonparametric counterpart of Pearson correlation coefficient), and chi-squared test.

In the sections below, we summarize the debates on how to analyze Likert-type data in the order of decision-making in applied statistics, robustness of statistical methods, and the underlying distribution of scores derived from Likert-type items. In our summary, we focus the discussion on group comparisons.

2.1. Decision-Making in Applied Statistics

Some researchers suggested that data analysis should not be restricted by the type of measurement (e.g., Lord, 1953; Tukey, 1986, Velleman & Wilkinson, 1993). Velleman and Wilkinson (1993) argued that the decision-making for choosing an appropriate statistical method should be guided by "the questions being investigated, the patterns discovered in the course of the analysis, and the additional data that may be available" (p. 71). Lord (1953) used a story of Professor X who sold football jersey numbers to college students from a vending machine to illustrate that appropriate statistical tests depend on the problem at hand not on the type of measurement. In the story, Professor X was said to be feeling guilty of computing means and standard deviations on ordinal numbers and he taught his students very carefully to adhere to Steven's theory of measurement. Therefore, when it was suspected that freshman team had lower jersey numbers than the sophomore team, Professor X had to consult with a statistician to understand whether freshmen had gotten low numbers just by chance. The statistician performed a parametric test on football numbers to determine whether a sample from the machine should be considered non-random. Although Lord's illustration received criticisms, it is generally agreed that measurement can be much complicated than it seems (Scholten & Borsboom, 2009). Using another example of raffle tickets, Velleman and Wilkinson (1993) explained that type of data is rarely fixed. The type of data depends on its interpretation and what additional information is available. When consecutively numbered raffle tickets are given to people, starting with 1, in the order that people enter a door for attending to an event, the number on the ticket may be interpreted as nominal data, ordinal data, interval data, or ratio data.

To deal with real-life data, transformations may be used to alternate certain characteristics of the scores for good data analysis. For instance, Zimmerman (1995) demonstrated when there are outliers, scores can be transformed to ranks before performing the *t* or *F* test. According to Kirk (2013), transformations are used to (1) achieve homogeneity of error variances, (2) achieve normality of error effects, (3) minimize the effects of extreme scores, and (4) obtain additivity of effects. Several transformations have been suggested in the literature, including the square-root transformation, logarithmic transformation, and rank-transformation (Box

et al., 2005; Hora & Iman, 1988; Kirk, 2013; Tukey, 1957; Zimmerman, 2011). These transformations are monotone but nonlinear, which means the transformations change the shape of the distribution but preserve order. Although studies have suggested transformations as strategies to analyze data not meeting the assumptions of statistical tests, such transformations are allowed for nominal and ordinal data only, based on Stevens (1946). According to Stevens, performing parametric tests (e.g., *t* or *F* tests) on rank data is not appropriate either.

While some researchers suggest that good data analysis does not assume data types, others examine the robustness of statistical methods to inform the decision of using parametric or nonparametric methods for analyzing Likert-type data.

2.2. Robustness of Statistical Methods

Robustness in statistics means "statistical methods which are relatively insensitive to: departure from distributional assumptions, outliers, sample censoring or other modifications, or large sample requirements" (Launer & Wilkinson, 1979, p. ix). To understand the robustness of a statistical method, simulation studies can be used to evaluate the performance of statistical methods in certain scenarios (e.g., population distribution is nonnormal). Simulation studies involve generating data by pseudo-random sampling from known distributions. The data may be generated by repeated sampling with replacement from a specific dataset or a known model (e.g., a standard normal distribution) once or many times (Morris et al., 2019). Using simulation studies from which data are generated by known models, researchers are able to manipulate the population. Thus, contrary to empirical research, what is true is known. For instance, when comparing the means of two independent groups, a researcher who generates the data from known models may manipulate whether there is a difference between the two population means or not. When there is no population mean difference and $n_{simulated}$ (e.g., 10,000) datasets are generated from the two populations, a preferred statistical method should yield results of rejecting the null hypothesis about α (the preselected acceptable probability of rejecting a true null hypothesis, also called nominal α) $\times 100$ percent of times of the simulated datasets. When the two population means are different, a preferred statistical method should yield the greatest number of rejecting the null hypothesis from the simulated datasets among all the alternatives. Zimmerman (1995) suggested "the probability distribution of a random variable, not the level of measurement, is paramount in determining which statistical test is appropriate" (p. 93).

Below we synthesize simulation results from comparing two or more groups using a single Likert-type item and total scores from many Likert-type items.

2.2.1. Using Single Likert-Type Item for Group Comparison

As early as 1969, Hsu and Feldt (1969) examined the performance of the *F* test when data were drawn from items with two to five response options. They manipulated the population distributions to be either symmetrical or moderately skewed (skewness ranged from 0 to 1.15). In some conditions, one population had a variance two times greater than the other(s). Hsu and Feldt included only equal sample size conditions and there were either 11 or 51 in each of the two or four groups. Results from Hsu and Feldt showed that Type I error rates were acceptable under all conditions for three to five response options, even when scores

were drawn from populations with unequal variances. When two response options were used, the Type I error control was the worst. In order to understand if the chi-squared test could be used as an alternative to the *F* test in detecting mean differences, Hsu and Feldt compared the results of the chi-squared test to the *F* test in some of their manipulated conditions. Findings showed that when there were a small number of participants in each group (e.g., 11 per group with five response options), the *F* test should be used instead of the chi-squared test.

Later, Nanna and Sawilowsky (1998) conducted a simulation study using data generated by sampling with replacement from empirical data collected from seven Likert-type items with seven response options and from the total score of the seven items. The empirical data were scores on items of the Functional Independence Measure obtained from patients when they were admitted to, and discharged from a rehabilitation hospital. In general, the distributions used in Nanna and Sawilowsky's study were more skewed than those in Hsu and Feldt (1969). Similar to Hsu and Feldt's study, Nanna and Sawilowsky manipulated only equal sample size conditions, and the sample sizes were 10, 20, 30, 40, and 60 for each of both groups. They compared the statistical power of the *t* test and Wilcoxon rank sum test. Nanna and Sawilowsky concluded that the Wilcoxon test outperformed the *t* test for almost all the manipulated conditions. It is worth noting that the results showed power advantages of the Wilcoxon rank sum test not only for single Likert-type items but also for the total scores from the seven Likert-type items. In addition, the power advantages of the Wilcoxon test over *t* test were held regardless of the sample size.

More recently, de Winter and Dodou (2010) compared the performance of the *t* test and Mann-Whitney *U* test for data obtained from Likert-type items with five response options. They manipulated 14 distributions and then generated data drawn from each pair of the 14 distributions (a total of 91 combinations) as well as data drawn from the same distribution. Three equal sample size conditions were used, including 10, 30, and 200 for each of the two groups. Two unequal sample size conditions were manipulated as 5 and 20 and 100 and 10. When data were drawn from one of the 91 combinations of distributions, de Winter and Dodou compared the Type II error rates of the *t* test and Mann-Whitney *U* test. When data of two groups were drawn from the same distribution, the Type I error rates of the two methods were compared and the nominal Type I error rate was set as 5%.

In terms of the Type I error rate, findings from de Winter and Dodou's study (2010) revealed that the two methods had the largest Type I error rate (7.4% for the *t* test and 7.7% for the Mann-Whitney *U* test, meaning the tests rejected true null hypothesis too many times) across all the manipulated conditions, when unequal sample sizes of 5 and 20 were drawn from the population distribution of very strongly agree (i.e., skewness = −3.70, kurtosis = 17.03, 0% of people responded 1 as *strongly disagree*, 1% responded 2 as *disagree*, 3% responded 3 as *neutral*, 6% responded 4 as *agree* and 90% responded 5 as *strongly agree*). In terms of Type II error rate, when sample sizes were both 10, most results showed either equal Type II error rates or smaller Type II error rates for the Mann-Whitney *U* test, except when one group was drawn from a strong multimodal distribution (i.e., skewness = 0, kurtosis = 1.06, 45% of people responded 1 as *strongly disagree*, 5% responded 2 as *disagree*, 0% responded 3 as *neutral*, 5% responded 4 as *agree* and 45% responded 5 as *strongly agree*). As sample sizes increased

from 10 to 200 in both groups, the number of conditions of no Type II error rate difference increased. Yet, when there were differences, the differences increased as the sample sizes increased. For example, when one group was drawn from a multimodal distribution (i.e., skewness = −0.83, kurtosis = 2.37, 15% of people responded 1 as *strongly disagree*, 5% responded 2 as *disagree*, 15% responded 3 as *neutral*, 25% responded 4 as *agree*, and 40% responded 5 as *strongly agree*), the maximum difference in Type II error rates between the two methods increased from 6%, 19%, to 62% as sample size increased from 10, 30, to 200, with the Mann-Whitney *U* yielding smaller Type II error rates. When one group was drawn from a strong multimodal distribution, the maximum difference in Type II error rates between the two methods increased from 21%, 26%, to 57% as sample size increased from 10, 30, to 200, with the *t* test yielding smaller Type II error rates. When sample sizes were unequal, the paring of groups with different sizes to the population distributions mattered. For example, when comparing a larger size group (n_1 = 100) drawn from a neutral peak distribution (i.e., skewness = 0.51, kurtosis = 2.68, 0 people responded 1 as *strongly disagree*, 20% responded 2 as *disagree*, 50% responded 3 as *neutral*, 20% responded 4 as *agree*, and 10% responded 5 as *strongly agree*) with a smaller size group (n_2 = 10) drawn from a multimodal distribution, the Type II error rate was the same in both methods. In contrary, when comparing a larger size group (n_1 = 100) drawn from a multimodal distribution with a smaller size group (n_2 = 10) drawn from a neutral peak distribution, the Type II error rate was smaller for the Mann-Whitney *U* test than for the *t* test (difference = 24%).

2.2.2. Using Total Scores from Several Likert-Type Items for Group Comparison

When total scores are obtained from several Likert-type items, simulation studies that generate data from different theoretical distributions can be used as references. In general, studies have shown that when populations have normal or uniform distributions, the *t* test has power advantages over the Mann-Whitney *U* test, regardless of the sample size (Boneau, 1962; Posten, 1984; Poncet et al., 2016). The Mann-Whitney *U* test has power advantage over the *t* test when population distributions are heavy tailed, such as lognormal, mixed normal, and chi-squared distributions (Nanna & Sawilowsky, 1998; Bridge & Sawilowsky, 1999; Poncet et al., 2016; Zumbo & Zimmerman, 1993).

It should be noted that neither the *t* test nor Mann-Whitney *U* test is robust to unequal variances—one of the assumptions for comparing two independent groups (Grissom, 2000; Nachar, 2008; Neuhäuser & Ruxton, 2009; Skovlund & Fenstad, 2001; Zumbo & Zimmerman, 1993). A solution for dealing with unequal variances is to use the Welch *t* test. The Welch *t* test was developed independently by Welch (1938) and Satterthwaite (1946). The *t* test (or Student *t* test/ independent-samples *t* test) assumes equal variances and uses the pooled variance in the denominator of the test static:

$$\text{Student } t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}, \qquad (1)$$

where $\bar{X}_1$ is the sample mean of the first group, $\bar{X}_2$ is the sample mean of the second group, s_p^2 is the pooled variance of the two groups, and n_1 and n_2 are the sample sizes of the first and second group, respectively. The calculation of s_p^2 is as follows:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}, \qquad (2)$$

where s_1^2 and s_2^2 are the variances of the first and second groups, respectively. The

Welch *t* test defines the test statistic as follows:

$$\text{Welch } t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \quad (3)$$

Welch *t* is approximately *t* distributed with df_{Welch}:

$$df_{\text{Welch}} = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}. \quad (4)$$

Because simulation studies showed that the Welch *t* test maintained appropriate Type I error rates and statistical power under equal and unequal variances conditions, literature recommended the use of the Welch *t* test for comparing two independent groups for a general purpose (Best & Rayner, 1987; Delacre et al., 2017; Fagerland & Sandvik, 2009; Rasch et al., 2011; Roxton, 2006; Skovlund & Fenstad, 2001; Zumbo & Zimmerman, 1993). The Welch *t* test is available in many statistical software packages. In R, the function t.test performs the Welch *t* test by default. In SPSS, T-TEST procedure can be used to perform both the Student *t* test and Welch *t* test. Results of the Student *t* test (equal variances assumed) and Welch *t* test (equal variances not assumed) are presented simultaneously in the table of Independent Samples Test table. T-TEST procedure is also accessible in the menus via Analyze>Compare Means>Independent-samples T Test. In SAS, PROC TTEST command yields the results of the Student *t* test (pooled) and Welch *t* test (Satterthwaite) and they are presented simultaneously in the output table.

A related simulation study conducted by Zimmerman (2011) showed power advantages of the Mann-Whitney *U* test over the *t* test for data generated from exponential, mixed-normal, lognormal, extreme value, half-normal and chi-squared distributions. Interestingly, when rank-transformation was applied for scores obtained from these nonnormal distributions and then the *t* test was applied to these transformed scores, the *t* test with rank-transformation scores had power advantages over the Mann-Whitney *U* test with the original data. Indeed, transformations of data may improve the performance of parametric methods. However, when interpreting results from tests on transformed data, one needs to be careful that the results are valid only on the transformed scale (Fagerland & Sandvik, 2009).

2.2.3. Tests of Statistical Assumption

Some researchers recommended performing tests of statistical assumption before employing a significance test for the null hypothesis of interest (Keppel, 1991; Keselman et al., 2014; Kirk, 2013; Lix & Keselman, 2004; Schoder et al., 2006; Triola et al. 2002). Before employing a *t* test, a normality test, such as the Shapiro-Wilk test, Kolmogorov-Smirnov test, and Anderson-Darling test, may be used to examine the normality assumption for the two groups. An equal variance test, such as Levene's test and *F*-ratio test, may be used to examine the equal variance assumption. However, not all tests for statistical assumptions perform equally and each test has its limitations.

Among various tests that can be used for testing normality, Keselman et al. (2014) reported that the performance of Anderson-Darling and Cramer-von Mises tests were acceptable when sample size was less than 100, including Likert-type data with five response options. Keselman also recommended Hochberg's sequentially-rejective Bonferroni procedure (Hochberg & Tamhane, 1987) for overall Type I error control (e.g., .15 or .20) of multiple normality tests. Users can use SAS to perform the Anderson-Darling and Cramer-von Mises tests. Levene's test may be used to test equal variance assumption. Yet, Levene's test has low statistical power when

sample sizes are small and unequal (Delacre et al., 2017; Nordstokke & Zumbo, 2007). In addition, Levene's test is too liberal when the populations are nonnormal (Conover et al., 1981). Several studies support the argument that it is unnecessary to perform the equal variance test for two group comparisons (Gans, 1981; Hayes & Cai, 2007; Moser & Stevens, 1992; Rasch et al., 2011).

2.3. Underlying Continuous Distributions for Likert-Type Data

2.3.1. Ordered Logit Model or Ordered Probit Model

There have been discussions on using ordered models to analyze Likert-type data, such as the ordered logit model or ordered probit model (Agresti, 2013; Becker & Kennedy, 1992; Daykin & Moffatt, 2002; Fielding, 1999; Greene & Hensher, 2010; Hoffmann, 2016; Liddell & Kruschke, 2018; Verhulst & Neale, 2021). In most cases, the logit model yields similar results to the probit model (Agresti, 2007; Hoffmann, 2016; Liddell & Kruschke, 2018). We therefore focus our discussion on the ordered probit model. The central idea for using the ordered probit models is that the ordered response is simply a set of discrete outcomes that by some criteria can be ordered. Furthermore, underlying the observed response is a latent, continuously distributed random outcome. For example, let y be the observed response to a Likert-type item with five response options (e.g., 1 = strongly disagree to 5 = strongly agree) and y can be 1, 2, 3, 4, or 5. Let y^* be the underlying latent outcome representing the propensity of individuals to agree with the statement of an item and X be the group membership (e.g., male or female). The basic ordered pobit model can be written as follows:

$$y^* = X\beta + \varepsilon, \tag{5}$$

where β is the coefficient and ε is the error term and it is assumed to be normally distributed with the mean of 0 and standard deviation of 1. The relationship between y^* and y is:

$$\begin{aligned} &y = 1 \text{ if } -\infty < y^* \le \tau_1 \\ &y = 2 \text{ if } \tau_1 < y^* \le \tau_2 \\ &y = 3 \text{ if } \tau_2 < y^* \le \tau_3 \\ &y = 4 \text{ if } \tau_3 < y^* \le \tau_4 \\ &y = 5 \text{ if } \tau_4 < y^* \le \infty \end{aligned} \tag{6}$$

In Equation 6, τ_1 to τ_4 are the threshold parameters and also known as cut points. If y^* falls into category j (= 1, 2, 3, 4, or 5), the observed response y is j. The model contains the β and the four threshold parameters to be estimated using the observed responses. Assuming the underlying distribution of the response is normally distributed, the probability associated with the observed response y is:

$$\text{Prob}(y) = \text{Prob}(\tau_{y-1} < y^* < \tau_y) = \Phi(\tau_y - X\beta) - \Phi(\tau_{y-1} - X\beta), \quad y = 1, 2, 3, \ldots J \tag{7}$$

where $\Phi(.)$ is the standard normal cumulative distribution function and J is the number of categories. Estimations of β, τ_1, τ_2, τ_3, and τ_4 can be done in maximum likelihood estimation.

When there is only one predictor for the ordered probit model, such as group membership, the significant test of the β reveals whether or not the group membership can be used to predict the response on the item. Multiple predictors can be included in the ordered probit model to predict the response on a Likert-type item. Using a Bayesian approach to analyze Likert-type data with five response options, Liddell and Kruschke (2018) found that the ordered probit model better described the data than the parametric method for both single Likert-type items and the means of multiple Likert-type items.

In R, the ordinal probit regression can be carried out in the MASS package under the polr function. In SPSS, GENLIN

procedure can be used to perform ordinal probit regression. It can also be accessed in the menus via Analyze>Generalized Linear Models>Generalized Linear Models. After this, one needs to choose "Ordinal probit" for Ordinal Response from the tab Type of Model. In SAS, the PROC LOGISTIC statement with the LINK = PROBIT can be used to fit data with a probit model.

2.3.2. Item Response Theory Model

Literature has suggested rescaling ordinal data to an interval scale using item response theory (IRT) model (Harwell & Gatti, 2001; Oon & Fan, 2017; Zhao et al., 2017). IRT is also known as latent trait theory. In IRT models, a person's score is quantified by the latent trait estimate, which is estimated on an interval scale. If the assumptions required for the applied IRT model are met, standard statistical procedures can be used to analyze the estimated underlying latent trait. IRT is usually compared to the classical test theory (CTT). CTT introduces three concepts: (observed) test score (X), true score (T), and error score (E). The fundamental model for CTT is

$$X = T + E. \quad (8)$$

In the book *Standards for Educational and Psychological Testing*, CTT is defined as "a psychometric theory based on the view that an individual's observed score on a test is the sum of a true score component for the test taker and an independent random error component" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 216). To solve Equation 8, assumptions are made in CTT models. For example, one assumption is that the average error score in the population of examinees is zero. Advantages of classical test models are the assumptions for CTT models, which are easy to meet in real test data. However, both person parameters (i.e., T) and item parameters (i.e., item difficulty and item discrimination) are dependent on the test and on the sample (Crocker & Algina, 2008; Hambleton & Jones, 1993).

In the book *Standards for Educational and Psychological Testing*, IRT is defined as "a mathematical model of the functional relationship between performance on a test item, the test item's characteristics, and the test taker's standing on the construct being measured." (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 220). In IRT, the logit model is usually used to link an individual's response to an item and his/her true latent trait. The simplest IRT model for items with two score categories (e.g., yes or no, correct or incorrect) is the one-parameter Rasch model. Rasch model uses a single difficulty parameter for each item and it assumes equal discrimination across items. The most frequently used models for polytomous items (e.g., Likert-type items) are the graded response, partial credit, and generalized partial credit models (Bandalos, 2018).

Advantages of IRT models are that(1) item characteristics (e.g., item difficulty) are independent of the individuals from which they were estimated, (2) the underlying latent trait level for each individual is estimated based on ones' response to each item by accounting for the characteristics of each item, and (3) item information (i.e., a function of the change in probability of a response) and its associated standard errors vary along the latent trait continuum (Crocker & Algina, 2008; Hambleton & Jones, 1993; Zhao et al., 2017). In CTT, each item on a scale is weighted equally for the total score. Therefore, it is assumed that the latent trait level required

to answer each item correctly (or to select a response) is the same across all the items. Unlike CTT, IRT treats the item difficulty as information to be incorporated in scaling items. There are two key assumptions of IRT models, namely unidimensionality and local independence. Unidimensionality means that a single latent variable accounts for variation common to items whereas local independence means that item responses are independent of one another after controlling for the ability or trait being measured (Bandalos, 2018; Crocker & Algina, 2008; Hambleton & Jones, 1993).

In the previous section, we introduce the ordered logit and probit model without random effects. When we estimate the two models with random effects, the ordered logit, ordered probit, and the IRT models can all be formulated within the generalized linear mixed model family (Greene & Hensher, 2010; DiTrapani et al., 2018). There are many packages within R for employing IRT models, such as ltm. SPSS does not have any built-in procedures for IRT models. The SPIRIT Macro in SPSS allows users to conduct one-parameter IRT for dichotomous or polytomous (applications of item response trees) response variables through the typical SPSS point-and-click (DiTrapani et al., 2018). The SPIRIT Macro can be downloaded at https://njrockwood.com/spirit. In SAS, users can use the PROC IRT statement to carry out IRT analyses. There are also software specifically designed for IRT, such as WINSTEPS, IRTPRO, BILOG-MG, and RUMM. Oon and Fan (2017) demonstrated one-parameter IRT analyses in WINSTEPS along with conducting parametric statistical tests based on the latent trait estimates. In a simulation study, Xu and Stone (2012) compared the results of using IRT trait estimates and CTT-based total scores in predicting outcomes. They concluded that results of the IRT trait estimates and CTT-based total scores were comparable in terms of predicting outcomes. Furthermore, they suggested that CTT-based total scores may outperform IRT trait estimates for scales of short length (10 items), especially when the sample size is small (N = 250).

3. Prevalence of Likert-Type Data in Educational Technology Research

To understand the prevalence of Likert-type data in the field of educational technology, we reviewed 424 articles that were published in *Educational Technology Research and Development* (ETR&D) over the most recent five-year period from 2016 to 2020. The review period of five years was recommended by Goodwin and Goodwin (1985) to detect a stable trend in research methodology. ETR&D is affiliated with the Association for Educational Communication and Technology. ETR&D is listed as one of the top 10 journals in the field of educational technology in Google Scholar with the h5-index of 41. Based on Journal Citation Reports, the 2019 Journal Impact Factor of ETR&D is 2.303.

There are three sections of ETR&D: the research section, the development section, and the cultural and regional perspectives section. The research section publishes studies on topics relating to applications of technology or instructional design. The development section publishes research on planning, implementation, evaluation and management of instructional technologies and learning environments. The cultural and regional perspectives section publishes research that is focused on how technologies are used to enhance learning, instruction, and performance specific to a culture or region. Our review included articles published in all the three sections and in special issues. We excluded errata, corrections, and awards announcements.

3.1. Research Question

The review on the 424 articles was guided by the following research questions:

1. What percentage of articles that were published in ETR&D between 2016 and 2020 used Likert-type data in answering research questions?
2. Among the reviewed articles that used Likert-type data as dependent variables, what were the strategies employed to analyze Likert-type dependent variables?
3. How did the authors of our reviewed articles deal with the assumptions of statistical method for analyzing Likert-type dependent variables?

3.2. Coding and Data Analysis

To answer Research Question 1, each article was coded into one of the five types: (1) Likert-type data as dependent variable(s) in inferential statistics to answer research questions, (2) Likert-type data as independent variable(s) in inferential statistics to answer research questions, (3) the main purpose of the article was to develop an instrument using Likert-type items, (4) responses to Likert-type items used in descriptive statistics only, and (5) no use of Likert-type data. Each article was reviewed in the order of the five types of articles, from the lower number to the higher number. When one article met the criteria for the lower number of the type of article, the article was coded as that category and would not be examined for meeting the criteria for the higher number of article type.

To answer Research Question 2, articles that were coded as using Likert-type data as dependent variable(s) were further examined for the strategies employed to perform statistical analysis. Each article was only coded into one type of the following statistical analyses: (1) ordered probit/logit model, (2) transformation before analyses, (3) Welch's t or F test, (4) using trait estimates from IRT model for analyses, (5) both parametric and nonparametric methods, (6) parametric method, and (7) nonparametric method. Similar to the coding mechanism we used to answer Research Question 1, when one article met the criteria for the lower number of the type of statistical analysis, the article was coded as that category and would not be examined for meeting the criteria for higher number of statistical analysis. Although these seven statistical analyses overlapped, the coding strategy helped us understand whether the analysis approaches in our literature review were actually used in educational technology research.

The articles used for answering Research Question 2 were again used for answering Research Question 3. Each article was coded as (1) making no mention of statistical assumption for dealing with Likert-type dependent variables, (2) describing terms related to statistical assumption (e.g., no outliers, reporting skewness and kurtosis within acceptable range), or (3) explicitly describing statistical tests performed for checking assumptions. We reported descriptive statistics to answer the three research questions.

3.3. Results

3.3.1. Results for Research Question One

Table 1 shows the frequencies of the five types of articles. In the last row of Table 1, the numbers in the parentheses are the percentages of the five types of articles across the five-year review period. About 52.4% of the reviewed articles did not use Likert-type data and 47.6%

of the articles did. Among the articles that used Likert-type data, the majority used Likert-type data as dependent variable(s) for inferential statistics (139 or 32.8%). Figure 1 presents the percentages of the five types of articles across the five- year review period. Across the five years, the use of Likert-type data in ETR&D articles was about 50%, more or less.

Table 1: Frequency of Likert-Type Data Used in ETR&D Articles

	Not used	Dependent variable	Independent variable	Instrument development	Descriptive only	Total
2016	34	18	1	1	7	61
2017	36	24	2	1	9	72
2018	31	28	2	3	8	72
2019	34	22	4	2	5	67
2020	87	47	5	2	11	152
2016-2020	222 (52.40%)	139 (32.80%)	14 (3.30%)	9 (2.10%)	40 (9.40%)	424 (100.0%)

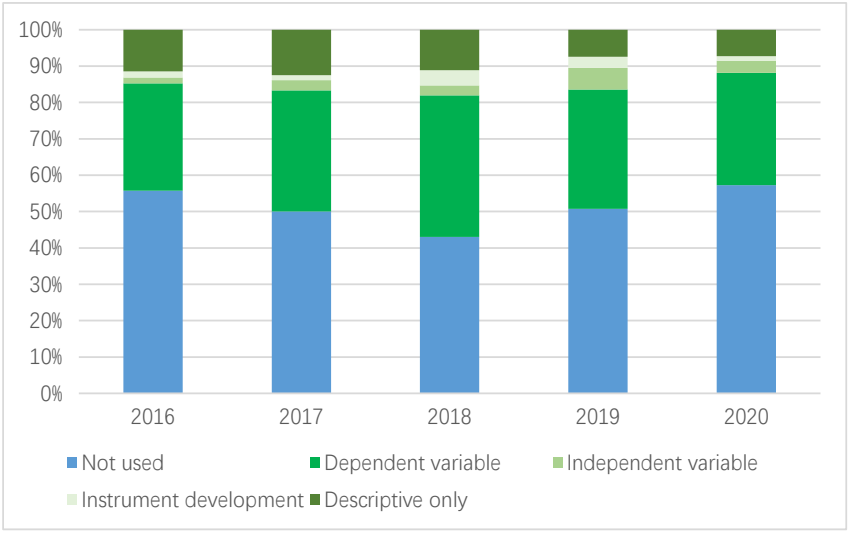

Figure 1: Percentage of Likert-Type Data Used in ETR&D Articles

3.3.2. Results for Research Question Two

Among the 139 articles that used Likert-type data as dependent variable(s), 2 (1.4%) articles employed an ordered logit/probit model, 1 (0.7%) article used transformation, 1 (0.7%) article applied the Welch's *F* test, 120 (86.3%) articles used parametric methods, 10 (7.2%) articles used nonparametric methods, and 5 (3.6%) articles used both parametric and nonparametric methods to deal with Likert-type dependent variables. We did not identify any articles that used trait estimates generated from IRT models to analyze Likert-type data.

3.3.3. Results for Research Question Three

As it is shown in Table 2, about half (72 or 51.8%) of the 139 articles that used Likert-type dependent variables did not mention statistical assumptions. Twenty-seven articles (19.4%) included statements related to statistical assumptions and 40 articles (28.8%) included statements about statistical assumptions being tested. Across the five-year review period, the first year (i.e., 2016) had the lowest percentage of no report of statistical assumptions and the last year (i.e., 2020) had the highest percentage of no report of statistical assumptions (Figure 2). We did not find any article used the Anderson-Darling or Cramer-von Mises test for normality assumption.

Table 2: Frequency of Statistical Assumption Reporting for Likert-Type Dependent Variable in ETR&D Articles

	No report	Assumption described	Assumption tested	Total
2016	7	7	4	18
2017	12	2	10	24
2018	15	2	11	28
2019	11	8	3	22
2020	27	8	12	47
2016-2020	72 (51.8%)	27 (19.4%)	40 (28.8%)	139 (100.0%)

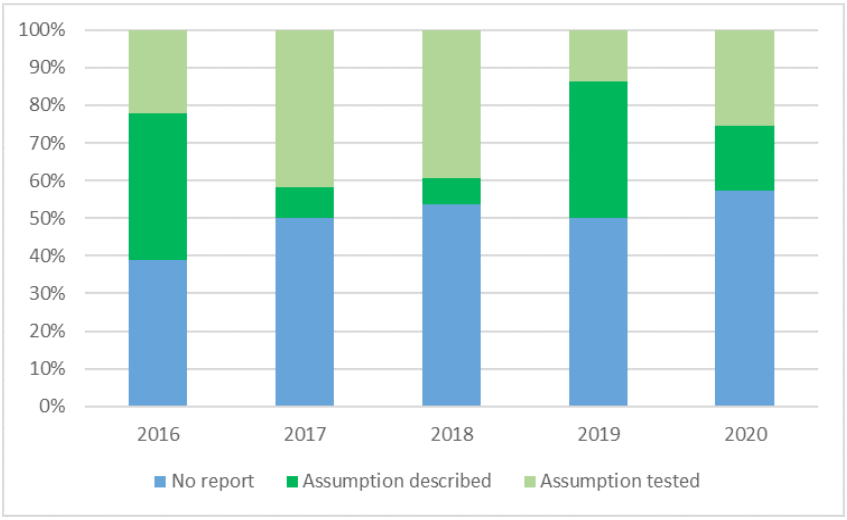

Figure 2: Percentage of Statistical Assumption Reporting for Likert-Type Dependent Variable in ETR&D Articles

4.Discussion and Conclusions

Likert-type items are widely used in education and related fields (Antonialli et al., 2017; Carifio & Perla, 2007; Edmondson, 2005; Harwell & Gatti, 2001; Liddell & Kruschke, 2018; Potvin & Hasni, 2014; Tsui, 1997). Yet there is no consensus among researchers regarding analysis strategies for handling Likert-type dependent variables (Carifio & Perla, 2007; Howell, 2013; Jamieson, 2004; Kuzon et al., 1996; Lord, 1953; Norman, 2010; Siegel, 1956; Sprinthall, 2012; Velleman & Wilkinson, 1993; Zimmerman, 2011). In this paper, we synthesized literature on the use and analysis of Likert-type data. To understand the prevalence of Likert-type data in educational technology research, we reviewed the 424 articles published in ETR&D from 2016 to 2020. In addition, we examined strategies that educational technology researchers employed to handle Likert-type dependent variables.

4.1. Comparing the Results of Current and Previous Studies

Findings from our review of ETR&D articles revealed that about 50% of the articles used Likert-type data. This number is lower than the percentage identified in Harwell and Gatti (2001). Harwell and Gatti (2001) reviewed articles published in the journals *American Educational Research Journal, Sociology of Education, and Journal of Educational Psychology* in 1997. Their findings concluded that 73% of the dependent variables used in these articles used Likert-type data. Reasons for the lower percentage in our study may relate to that alternative measurements have been used in educational technology research. It is also possible that over the 20 years, there have been more alternative measures developed for measuring attitudes and opinions.

From reviewing the literature, we grouped the discussion on handling Likert-type data into three categories: considering good decision-making in applied statistics, investigating robustness of the methods, and considering the underlying distribution of scores derived from Likert-type items. Based on our review, we identified multiple strategies for handling Likert-type data, including traditional parametric method (e.g., *t* test, *F* test), traditional nonparametric method (e.g., Mann-Whitney *U* test, Kruskal-Wallis test), transformation of Likert-type data to change the shape of score distribution before employing traditional parametric method, Welch *t* test, application of an ordered probit/logit model, and application of trait estimates generated from IRT model to rescale Likert-type data to interval data before employing the traditional parametric method. We also provided computing tools for conducting the Welch *t* test, ordered probit/logit model, and IRT model. The majority of our reviewed ETR&D articles (86.3%) employed traditional parametric methods (e.g., *t* test, *F* test) to deal with Likert-type dependent variables. In addition, less than 3% of ETR&D articles employed an ordered probit/logit model, transformation of scores, or IRT model to analyze Likert-type dependent variables. These findings were similar to Liddell and Kruschke (2018). Liddell and Kruschke reviewed 68 articles that were published in the journals *Journal of Personality and Social Psychology, Psychological Science, and Journal of Experimental Psychology: General.* They reported none of their reviewed articles that used Likert-type data as a dependent variable employed an ordinal model to analyze the data. We suggest that researchers may believe that parametric methods are robust to violation of statistical assumptions. The lower rate of using methods other than traditional parametric methods to handle Likert-type data may also relate to researchers' unfamiliarity

with the alternative methods.

4.2. Limitations and Suggestions

We acknowledge that our synthesis did not exhaust all the strategies proposed in literature to handle Likert-type data. For example, Camparo and Camparo (2013) proposed the state multipole method to analyze Likert-type data. Robust measures proposed by Wilcox (2017) may also be used to analyze data collected from Likert-type items. In addition, findings from our review of ETR&D articles may not be generalized to articles published in other educational technology journals. We hope that this paper not only provides a preliminary understanding of current practice in analyzing Likert-type data in educational technology but also empowers educational technology researchers to effectively analyze Likert-type data.

To inform educational technology researchers about alternative strategies for handling Likert-type data, we provide our suggestions and insight below:

- To compare two groups using Likert-type data for general purpose, researchers can use the Welch *t* test.
- If population variances are assumed to be equal, researchers can use the Mann-Whitney *U* test.
- If population variances are unequal and populations are not normally distributed, researchers may consider employ transformations before conducting parametric tests. When transformation is performed, the findings need to be interpreted using the transformed scale.
- When the underlying distribution for Likert-type data is assumed to be normal, researchers can use an ordered probit or logit model to analyze the data.
- To statistically test the normality assumption, researchers can use the Anderson-Darling or Cramer-von Mises test available in SAS. Although the Levene's test is the default for testing equal variance in SPSS, it may either have low statistical power or inflated Type I error rate.
- Approximately half of the ETR&D articles in our review did not mention statistical assumption. It brings concerns about whether data screening were performed. We would like to emphasize the importance of data screening to inform statistical analysis. The decisions on statistical analysis include, but are not limited to strategies to handle outliers and missing data and the choice of statistical test.
- When the Mann-Whitney *U* test is more powerful than the Student *t* test, the power advantage maintains regardless of sample size.
- The property of a statistical method remains the same for scores yielded by a single Likert-type item and for total or mean scores of several Likert-type items.

We hope the information presented in this paper could be of reference for educators and researchers who are interested in the relevant area of work. Research on appropriate statistical methods to analyze Likert-type data would be conducted further. Comments and suggestions are appreciated.

References

Agresti, A. (2007). An introduction to categorical data analysis (2nd ed.). John Wiley & Sons.

Agresti, A. (2013). Categorical data analysis (3rd ed.). John Wiley & Sons.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for educational and psychological testing. American Educational Research Association.

Antonialli, F., Antonialli, L. M., & Antonialli, R. (2017). Uses and abuses of the Likert scale: Bibliometric study in the proceedings of ENANPAD from 2010 to 2015. Reuna, 22(4), 1–19. http://dx.doi.org/10.21714/2179-8834/2017v22n4p1-19

Arbaugh, J. B., Cleveland-Innes, M., Diaz, S. R., Garrison, D. R., Ice, P., Richardson, J. C., & Swan, K. P. (2008). Developing a community of inquiry instrument: Testing a measure of the Community of Inquiry framework using a multi-institutional sample. The Internet and Higher Education, 11(3-4), 133–136. https://doi.org/10.1016/j.iheduc.2008.06.003

Areepattamannil, S, & Santos, I. M. (2019). Adolescent students' perceived information and communication technology (ICT) competence and autonomy: Examining links to dispositions toward science in 42 countries. Computers in Human Behavior, 98, 50–58. https://doi.org/10.1016/j.chb.2019.04.005

Bandalos, D. L. (2018). Measurement theory and applications for the social sciences. The Guilford Press.

Becker, W. E., & Kennedy, P. E. (1992). A graphical exposition of the ordered probit. Econometric Theory, 8(1), 127–131. https://doi.org/10.1017/S0266466600010781

Best, D. J., & Rayner, J. C. W. (1987). Welch's approximate solution for the Behrens–Fisher problem. Technometrics, 29(2), 205–210. https://doi.org/10.2307/1269775

Boneau, C. A. (1962). A comparison of the power of the U and *t* tests. Psychological Review, 69(3), 246–256. http://doi.org/10.1037/h0047269

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). Statistics for experimenters: Design, innovation, and discovery (2nd ed.). John Wiley.

Bridge, P. D., & Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: Comparative power of the *t*-test and Wilcoxon rank-sum test in small samples applied research. Journal of Clinical Epidemiology, 52(3), 229–235. https://doi.org/10.1016/S0895-4356(98)00168-1

Camparo, J., & Camparo, L. B. (2013). The analysis of Likert scales using state multipoles: An application of quantum methods to behavioral sciences data. Journal of Educational and Behavioral Sciences, 38(1), 81–101. https://doi.org/10.3102/1076998611431084

Carifio, J., & Perla, R. J. (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), 106–116. https://doi.org/10.3844/jssp.2007.106.116

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing

Likert scales. Medical Education, 42(12), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics, 23(4), 351–361. https://doi.org/10.2307/1268225

Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. CENGAGE Leaning.

Daykin, A. R., & Moffatt, P. G. (2002). Analyzing ordered responses: A review of the ordered probit model. Understanding Statistics, 1(3), 157–166. https://doi.org/10.1207/S15328031US0103_02

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. International Review of Social Psychology, 30(1), 92–101. https://doi.org/10.5334/irsp.82

de Winter, J. C. F., & Dodou, D. (2010). Five-point Likert items: *t* test versus Mann-Whitney-Wilcoxon. Practical Assessment, Research, and Evaluation, 15(11), 1–16. https://doi.org/10.7275/bj1p-ts64

DiTrapani, J., Rockwood, N., & Jeon, M. (2018). Explanatory IRT analysis using the SPIRIT macro in SPSS. The Quantitative Methods for Psychology, 14(2), 81–98. https://doi.org/10.20982/tqmp.14.2.p081

Edmondson, D. R. (2005). Likert scales: A history. In L. C. Neilson (Ed.), Proceedings of the 12th conference on historical analysis and research in marketing (CHARM) (pp.127–133).

Fagerland, M. W., & Sandvik, L. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. Contemporary Clinical Trials, 30(5), 490–496. https://doi.org/10.1016/j.cct.2009.06.007

Fielding, A. (1999). Why use arbitrary points scores?: ordered categories in models of educational progress. Journal of the Royal Statistical Society. Series A, Statistics in Society, 162(3), 303–328. https://doi.org/10.1111/1467-985X.00137

Gans, D. J. (1981). Use of a preliminary test in comparing two sample means. Communications in Statistics - Simulation and Computation, 10(2), 163–174.

Goodwin, L. D., & Goodwin, W. L. (1985). Statistical techniques in "AERJ" articles, 1979–1983: The preparation of graduate students to read the educational research literature. Educational Researcher, 14(2), 5–11. https://doi.org/10.3102/0013189X014002005

Greene, W. H., & Hensher, D. A. (2010). Modeling ordered choices: A primer. Cambridge University Press.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. Journal of Consulting and Clinical Psychology, 68(1), 155–165. https://doi. org/10.1037/0022-006X.68.1.155

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice, 12(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Harwell, M. R. & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. Review of Educational Research, 71(1), 105–131. https://doi.org/10.3102/00346543071001105

Hays, A. F., & Cai, L. (2007). Further evaluating the conditional decision rule for comparing two independent means. British Journal of Mathematical and Statistical Psychology, 60(2), 217–244. https://doi.org/10.1348/000711005X62576

Hochberg, Y. A., & Tamhane, A. C. (1987). Multiple comparison procedures. John Wiley & Sons.

Hoffmann, J. P. (2016). Regression models for categorical, count, and related variables: An applied approach. University of California Press.

Hora, S. C., & Iman, R. L. (1988). Asymptotic relative efficiencies of the rank-transformation procedure in randomized complete block designs. Journal of the American Statistical Association, 83(402), 462–470. https://doi.org/10.1080/01621459.1988.10478618

Howell, D. C. (2013). Statistical methods for psychology (8th ed.). Cengage.

Hsu, T.-C., & Feldt, L. S. (1969). The effect of limitations on the number of criterion score values on the significance level of the F-test. American Educational Research Journal, 6(4), 515–527. https://doi.org/10.2307/1162248

Jamieson, S. (2004). Likert scales: How to (ab)use them. Medical Education, 38(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Keppel, G. (1991). Design and analysis: A researcher's handbook (3rd ed.). Prentice Hall.

Keselman, H. J., Othman, A. R., & Wilcox, R. R. (2014). Testing for normality in the multi-group problems: Is this a good practice? Clinical Dermatology, 2(1), 29–43.

Kirk, R. E. (2013). Experimental design: Procedures for the behavioral sciences (4th ed.). Sage Publications.

Kuzon, W. M. Jr., Urbanchek, M. G., & McCabe, S. (1996). The seven deadly sins of statistical analysis. Annals of Plastic Surgery, 37(3), 265–272. https://doi.org/10.1097/00000637-199609000-00006

Launer, R. L., & Wilkinson, G. N. (Eds.) (1979). Robustness in statistics. Academic Press.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? Journal of Experimental Social Psychology, 79, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 22(140), 5–55.

Lix, L. M., & Keselman, H. J. (2004). Multivariate tests of means in independent groups designs. Effects of covariance heterogeneity and nonnormality. Evaluation & the Health Professions, 27(1), 45–69. https://doi.org/10.1177/0163278703261213

Lord, F. M. (1953). On the statistical treatment of football numbers. American Psychologist, 8(12), 750–751. https://doi.org/10.1037/h0063675

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38(11), 2074–2102. https://doi.org/10.1002/sim.8086

Moser, B. K., & Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test.American Statistician, 46(1), 19–21. https://doi.org/10.2307/2684403

Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. Tutorials in Quantitative Methods for Psychology, 4(1), 13–20. https://doi.org/10.20982/tqmp.04.1.p013

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. Psychological Methods, 3(1), 55–67. https://doi.org/10.1037/1082-989X.3.1.55

Neuhäuser, M. & Ruxton, G. D. (2009). Distribution-free two-sample comparisons in the case of heterogeneous variances. Behavioral Ecology and Sociobiology, 63, 617–623. https://doi.org/10.1007/s00265-008-0683-4

Nordstokke, D. W. and Zumbo, B. D. (2007). A cautionary tale about Levene's tests for equal variances. Journal of Educational Research & Policy Studies, 7(1), 1–14.

Norman, G. (2010). Likert scales, levels of measurement and the ''laws'' of statistics. Advances in Health Sciences Education, 15, 625–632. https://doi.org/10.1007/s10459-010-9222-y

Oon, P.-T. & Fan, X. (2017). Rasch analysis for psychometric improvement of science attitude rating scales. International Journal of Science Education, 39(6), 683–700. https://doi.org/10.1080/09500693.2017.1299951

Poncet, A., Courvoisier, D. S., Combescure, C., & Perneger, T. V. (2016). Normality and sample size do not matter for the selection of an appropriate statistical test for two-group comparisons. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 12(2), 61–71. https://doi.org/10.1027/1614-2241/a000110

Posten, H. O. (1984). Robustness of the two-sample *t*-test. In D. Rash & M. L. Tiku (Eds.), Robustness of statistical methods and nonparametric statistics (pp. 92–99). Springer.

Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at K-12 levels: A systematic review of 12 years of educational research. Studies in Science Education, 50(1), 85–129. https://doi.org/10.1080/03057267.2014.881626

Rasch, D., Kubinger, K. D. & Moder, K. (2011). The two-sample *t* test: pre-testing its assumptions does not pay off. Statistical Papers, 52, 219–231. https://doi.org/10.1007/s00362-009-0224-x

Roxton, G. D. (2006). The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann-Whitney *U* test. Behavioral Ecology, 17(4), 688–690. https://doi.org/10.1093/beheco/ark016

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin, 2(6), 110–114. https://doi.org/10.2307/3002019

Schmidt, D. A., Baran, E., Thompson, A. D., Mishra, P., Koehler, M. J., & Shin, T. S. (2009). Technological pedagogical content knowledge (TPACK): The development and validation of an assessment instrument for preservice teachers. Journal of Research on Technology in Education, 42(2), 123–149. https://doi.org/10.1080/15391523.2009.10782544

Schoder, V., Himmelmann, A., & Wilhelm, K. P. (2006). Preliminary testing for normality: Some statistical aspects of a common concept. Clinical Dermatology, 31(6), 757–761. https://doi.org/10.1111/j.1365-2230.2006.02206.x

Scholten, A. Z., & Borsboom, D. (2009). A reanalysis of Lord's statistical treatment of football numbers. Journal of Mathematical Psychology, 53(2), 69–75. http://dx.doi.org/10.1016/j.jmp.2009.01.002

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. McGraw-Hill.

Skovlund, E., & Fenstad, G. U. (2001). Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? Journal of Clinical Epidemiology, 54(1), 86–92. https://doi.org/10.1016/S0895-4356(00)00264-X

Sprinthall, R. C. (2012). Basic statistical analysis (9th ed.). Allyn & Bacon.

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Stevens, S. S. (1955). On the averaging of data. Science, 121(3135), 113–116. https://doi.org/10.1126/science.121.3135.113

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. Science, 161(3844), 849–856. https://doi.org/10.1126/science.161.3844.849

Triola, M. F., Goodman, W. M., & Law, R. (2002). Elementary statistics. Wesley.

Tsui, M.-S. (1997). Empirical research on social work supervision: The state of the art (1970–1995). Journal of Social Service Research, 23(2), 39–54. https://doi.org/10.1300/J079v23n02_03

Tukey, J. W. (1957). On the comparative anatomy of transformations. Annuals of Mathematical Statistics, 28(3), 602–632. https://doi.org/10.1214/aoms/1177706875

Tukey, J. W. (1986). Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning badmandments. In L. V. Jones (Ed.), The collected works of John W. Tukey. Volume III: Philosophy and principles of data analysis: 1949-1964 (pp. 187-389). Belmont, CA: Wadsworth.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. The American Statistician, 47(1), 65–72. https://doi.org/10.1080/00031305.1993.10475938

Verhulst, B., & Neale, M. C. (2021). Best practices for binary and ordinal data analyses. Behavior Genetics. Advance Online Publication. https://doi.org/10.1007/s10519-020-10031-x

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. Biometrika, 29(3/4), 350–362. https://doi.org/10.2307/2332010

Wilcox, R. R. (2017). Introduction to robust estimation and hypothesis testing (4th ed.). Academic Press.

Xu, T, & Stone, C. A. (2012). Using IRT trait estimates versus summated scores in predicting outcomes. Educational and Psychological Measurement, 72(3), 453–468. https://doi.org/10.1177/0013164411419846

Zhao, Y., Huen, J. M. Y., & Chan, Y. W. (2017). Measuring longitudinal gains in student learning: A comparison of Rasch scoring and summative scoring approaches. Research in Higher Education, 58, 605–616. https://doi.org/10.1007/s11162-016-9441-z

Zimmerman, D. W. (1995). Increasing the power of the ANOVA *F* test for outlier-prone distributions by modified ranking methods. Journal of General Psychology, 122(1), 83–94. https://doi.org/10.1080/00221309.1995.9921224

Zimmerman, D. W. (2011). Scales of measurement and choice of statistical methods. In M. Lovric (Ed.), International encyclopedia of statistical science (pp. 1285–1288). Springer.

Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? Canadian Psychology, 34(4), 390–400. https://doi.org/10.1037/h0078865

Contact the Author

Li-Ting Chen
Assistant Professor
Department of Educational Studies,
University of Nevada, Reno, USA.
Email: litingc@unr.edu

Leping Liu,
Professor,
Department of Educational Studies,
University of Nevada, Reno, USA.
Email: liu@unr.edu