Equidistance of Likert-Type Scales and Validation of Inferential Methods Using Experiments and Simulations

Bjorn Lantz
School of Engineering, University of Boras, Sweden
Bjorn.Lantz@hb.se

Abstract: Likert-type data are often assumed to be equidistant by applied researchers so that they can use parametric methods to analyse the data. Since the equidistance assumption rarely is tested, the validity of parametric analyses of Likert-type data is often unclear. This paper consists of two parts where we deal with this validity problem in two different respects. In the first part, we use an experimental design to show that the perceived distance between scale points on a regular five-point Likert-type scale depends on how the verbal anchors are used. Anchors only at the end points create a relatively larger perceived distance between points near the ends of the scale than in the middle (*end-of-scale effect*), while anchors at all points create a larger perceived distance between points in the middle of the scale (*middle-of-scale effect*). Hence, Likert-type scales are generally not perceived as equidistant by subjects. In the second part of the paper, we use Monte Carlo simulations to explore how parametric methods commonly used to compare means between several groups perform in terms of actual significance and power when data are assumed to be equidistant even though they are not. The results show that the preferred statistical method to analyse Likert-type data depends on the nature of their non-equidistance as well as their skewness. Under middle-of-scale effect, the omnibus one-way ANOVA works best when data are relatively symmetric. However, the Kruskal-Wallis test works better when data are skewed except when sample sizes are unequal, in which case the Brown-Forsythe test is better. Under end-of-scale effect, on the other hand, the Kruskal-Wallis test should be preferred in most cases when data are at most moderately skewed. When data are heavily skewed, ANOVA works best unless when sample sizes are unequal, in which case the Brown-Forsythe test should be preferred.

Keywords: Likert-type scale; equidistance; Monte Carlo simulation; ANOVA; Kruskal-Wallis test; Brown-Forsythe test; Welch test

1. Introduction

Since the psychologist Rensis Likert (Likert, 1932) published his seminal work on measurement of attitudes, Likert items (often referred to as 'Likert-type scales') and true Likert scales have been important data collection methodologies in research on attitudes and opinions in the social sciences in general. In business and management, Likert-type scales are often used by researchers to collect data (Alexandrov, 2010). Even though the optimal number of steps has been debated over the years (Pearse, 2011), the classical Likert item based on a statement where the subjects are asked to choose one out of five possible degrees of agreement, ranging from 'strongly agree' to 'strongly disagree', that complies with their view on the matter, still seem to the most common choice among researchers. According to the traditional classification of measurement scales (Stevens, 1946), Likert-type scales must be either ordinal or interval, depending on whether or not the scale is equidistant, since rank-ordering is possible while a true zero point is missing. Whether Likert-type scales should generally be regarded as ordinal or interval has been extensively debated among researchers over the years (e.g. Carifio and Perla, 2007; Jamieson, 2004; Michell, 1986), since the choice of statistical methodology depends on the result of this debate. According to the standard textbook view, data having only ordinal properties should be analysed with non-parametric statistics based on ranks. However, parametric statistics that are more powerful are allowed for data with interval properties, utilising the actual values of the data instead of just their ranks.

To be able to use parametric statistics on ordinal data, several different methods for 'rescaling' ordinal scales to get interval properties have been proposed (e.g. Granberg-Rademacker, 2010; Wu, 2007; King et al., 2004; Harwell and Gatti, 2001; Bendixen and Sandler, 1995). The use of such methods in applied research seems rare in practical analysis of Likert-type data, perhaps because it appears improbable that an ordinal scale with unknown distance between the scale points could actually be rescaled to a true interval scale, or due to a concern that the informational content of the data will be affected by the procedure. Instead, the vast majority of researchers about to analyse Likert-type data seem to either obey the purist view on statistics and open the non-parametric toolbox or trust the methodology researchers who claim that t and F are statistics robust to minor violations in the underlying assumptions. To be allowed to use parametric methods, many researchers simply assume that their scale has interval properties (Albaum, 1997) while a few put a lot of

Reference this paper as: Lant, B. "Equidistance of Likert-Type Scales and Validation of Inferential Methods Using Experiments and Simulations" *The Electronic Journal of Business Research Methods* Volume 11 Issue 1 2013 (pp 16-28), available online at www.ejbrm.com

effort into showing that their scale demonstrates acceptable quality as an interval measurement device (e.g. Albaum et al, 1977). The main problem, however, is that presence of interval properties is a necessary but not a sufficient condition for statistical analyses with parametric methodologies. Assumptions regarding normality and homoscedasticity must also be addressed.

There has been a considerable amount of research during the years testing the robustness of parametric methods like the t-test and the analysis of variance (ANOVA), often based on Monte Carlo approaches, against violations of the normality and the homoscedasticity assumptions. For example, Glass et al. (1972) found that skewness only had a small effect on the efficiency of ANOVA. Feir and Toothaker (1974) compared ANOVA with the Kruskal-Wallis test in a Monte Carlo study and concluded, based on 'the instability of power for the Kruskal-Wallis test', that ANOVA was the recommended procedure even when normality and/or homoscedasticity is doubtful. Zimmerman (1998) evaluated the Wilcoxon test against the t-test in a simulation study under concurrent violation of two assumptions, namely normality and homoscedasticity. He showed that nonparametric methods do not generally provide protection against concurrent violation of normality and homoscedasticity. Under some conditions, the Wilcoxon test would even make the situation worse. Lantz (2012) showed that parametric methods are generally more sensitive to different degrees of sample non-normality when populations are distinctly non-normal. He concluded that the Kruskal-Wallis test should be preferred as soon as the underlying populations are not known to be normal or approximately normal in order to avoid a preliminary test for normality that makes the overall level of significance unclear. However, despite extensive search, we have been unable to find any study that assesses the violation of the equidistance assumption while using parametric methods from a robustness perspective.

When researchers are in the process of choosing a statistical methodology to analyse Likert-type data, they should consider the way subjects perceive the response scale. If the scale is perceived as equidistant, parametric methods can obviously be used to analyse the data. If not, the purist view on statistics requires a non-parametric methodology. Research based on rescaling of ordinal data indicate that subjects actually do perceive Likert-type scales as non-equidistant, at least for specific constructs (e.g. Lee and Soutar, 2010; Mundy and Dickinson, 2004; Kennedy, Riquier, and Sharp, 1996; Bendixen and Sandler, 1995). Hence, the purpose of this study is to explore whether subjects generally perceive the five-point Likert-type scale as non-equidistant, and to examine how non-equidistance can affect the choice of statistical method for analysing Likert-type data.

The remainder of the paper is organised as follows. In the next section, two experiments are conducted in order to explore subject perceptions of the five-point Likert-type scale. Thereafter, a simulation study compares the performance of one-way ANOVA with alternative methods under different types of non-equidistance, followed by the conclusion.

2. Experiments

There are several obvious problems related to the process of assigning values to the points (or the distances between points) on a scale that is qualitative rather than quantitative by nature. Both the subjective distances between the scale points and the subjective zero points may differ between occasions and/or between respondents. They may even shift within the occasion due to respondent stimulus. In addition, the subjective distances between scale points may differ between subsets of the scale. Because of these problems, it is meaningless to assume that the distance between different points on the scale is measurable in absolute terms in order to be able to make a claim about the absolute distance between the points on the scale. In general, if the premise (e.g. it is actually possible to measure the absolute distance between points on a scale) cannot be proven valid, the result (e.g. measured absolute distances between points on the scale) may be invalid. For this reason, we did not assume that the absolute distances between the scale point is possible to measure. Instead, we only assumed that respondents can compare pair-wise changes in opinion represented by different movements between points on a five-point response scale within a Likert item, and express whether they perceive one change in opinion as greater than the other, or if they view them as equal.

The response scale formats under study here were technically five-point discrete visual analogue scales. Verbal anchors were used at the ends as illustrated in figure 1a in the first experiment, and at all points as illustrated in figure 1b in the second experiment, in order to examine whether respondents perceive the scale differently when they have to attach their personal interpretation of the different scale points based on only the numbers identifying them. Providing consecutive integers for the scale points was also assumed to maximize the likelihood that respondents actually would perceive the scale as interval, which meant that we would have a stronger case if the interval hypothesis were rejected.

	Strongly disagree				Strongly agree
Statement	1	2	3	4	5

Figure 1a: Verbal anchors only at the end points

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Statement	1	2	3	4	5

Figure 1b: Verbal anchors at all points

Two convenience samples of 178 and 101 respondents were recruited for participation in the experiments. No individual demographic or other background data were collected; however, all respondents were regular first- or second-year business students with an approximately even distribution between males and females. A majority of respondents were in their twenties. Hence, as always in experiments where students are used as subjects, the validity of the results rests on the assumption that students are representative of 'real people' (e.g., Cunningham et al., 1974). All respondents answered the questionnaire anonymously. The instructions specified that there was no 'correct' answer to any of the problems and that the aim of the study was to explore how people perceive this kind of response scale. All respondents were presented with a questionnaire where the scale was displayed and followed by nine different problems of the type illustrated in figure 2.

Compare a movement from 1 to 2 on the scale with a movement from 3 to 4 on the scale. Which one of them do you think represents a greater change in the level of agreement? Tick the appropriate box.

□ 1 to 2 is greater
□ 3 to 4 is greater
□ No difference

Figure 2: Example of a problem

To ensure reliability, several versions of the questionnaire were used so that different respondents were exposed to the problems and the options in different orders using a randomized process. The direction of the response scale was also varied randomly. No significant differences were found between the groups of respondents using different versions of the questionnaire.

The data from the problems in both experiments were analysed with respect to two different null hypotheses:

H1: A majority of the subjects do not perceive a difference.

H2: Uniform distribution characterises the two types of perceived differences.

The results from experiment 1, where only the end points on the scale had verbal anchors, are presented in table 1. Note that, for clarity, the nine comparisons are presented in four groups with different characteristics.

Group I consists of the three symmetric problems. In group II, there are two asymmetric problems with one-point movements on each side of the scale. Group III consists of the two asymmetric problems with two-point movements on each side of the scale. Finally, in group IV, we have the two possible one-point movements on the same side of the scale.

Table 1: Results from experiment 1 with verbal anchors only at the end points

Group	Movement comparison		Frequency (percentage)			H1	H2	
	A	B	A is greater	B is greater	No difference	p-value*	p-value*	Effect size h
I	1 to 2	4 to 5	43 (24.2%)	40 (22.5%)	95 (53.4%)	0.815	0.742	
	1 to 3	3 to 5	36 (20.2%)	40 (22.5%)	102 (57.3%)	0.973	0.647	
	2 to 3	3 to 4	29 (16.3%)	26 (14.6%)	113 (63.5%)	0.999	0.686	
II	1 to 2	3 to 4	79 (44.4%)	49 (27.5%)	50 (28.1%)	<0.001	0.010	0.47
	2 to 3	4 to 5	52 (29.2%)	78 (43.8%)	48 (27.0%)	<0.001	0.025	0.40
III	1 to 3	2 to 4	85 (47.8%)	52 (29.2%)	41 (23.0%)	<0.001	0.006	0.49
	2 to 4	3 to 5	50 (28.1%)	79 (44.4%)	49 (27.5%)	<0.001	0.013	0.45
IV	1 to 2	2 to 3	84 (47.2%)	43 (24.2%)	51 (28.7%)	<0.001	0.001	0.66
	3 to 4	4 to 5	51 (28.7%)	78 (43.8%)	49 (27.5%)	<0.001	0.020	0.42
* Based on regular two-sample z-tests								

For all three problems in group I, the symmetric comparisons, the results indicate that a majority of subjects do not perceive a difference. Hence, they experience the scale as symmetrical, which we refer to as the *symmetry effect*. For group II with the asymmetric one-step comparisons, a majority of the subjects do perceive differences. Specifically, significantly more subjects perceive 1–2 as larger than 3–4. We see the same effect on the other side of the scale where significantly more subjects perceive 4–5 as larger than 2–3. Besides the symmetry effect, these results also indicate that respondents tend to think that the distance between scale points is greater near an end of the scale than near the middle. We will refer to this as the *end-of-scale effect*. For the asymmetric two-step comparisons in group III, the same pattern emerges as in group II. A majority of the subjects perceive differences, and significantly more subjects perceive 1–3 as larger than 2–4, while significantly more subjects perceive 3–5 as larger than 2–4. Thus, both the end-of-scale effect and the symmetry effect can be seen in these data. Finally, in group IV, a majority of the subjects perceive differences while both the end-of-scale effect and the symmetry effect are visible. On the lower part of the scale, significantly more subjects perceive 1–2 as larger than 2–3. Mirroring this, significantly more subjects perceive 4–5 as larger than 3–4. Again, the distance between points near the end of the scale is experienced as greater changes in opinion than between points in the middle of the scale. In addition, this effect exists in a symmetrical manner on both sides of the scale.

Table 2 presents the results from experiment 2, where all points on the scale had verbal anchors. Again, the nine comparisons are divided into four groups with different characteristics.

Table 2: Results from experiment 2 with verbal anchors at all points

Group	Movement comparison		Frequency (percentage)			H1	H2	
	A	B	A is greater	B is greater	No difference	p-value*	p-value*	Effect size h
I	1 to 2	4 to 5	19 (18,8 %)	27 (26,7 %)	55 (54,5 %)	0.814		
	1 to 3	3 to 5	16 (15,8 %)	30 (29,7 %)	55 (54,5 %)	0.814		
	2 to 3	3 to 4	19 (18,8 %)	18 (17,8 %)	64 (63,4 %)	0.995		
II	1 to 2	3 to 4	24 (23,8 %)	62 (61,4 %)	15 (14,9 %)	<0.001	<0.001	0.92
	2 to 3	4 to 5	68 (67,3 %)	24 (23,8 %)	9 (8,9 %)	<0.001	<0.001	1.00
III	1 to 3	2 to 4	23 (22,8 %)	65 (64,4 %)	13 (12,9 %)	<0.001	<0.001	1.00
	2 to 4	3 to 5	51 (50,5 %)	32 (31,7 %)	18 (17,8 %)	<0.001	0.042	0.46
IV	1 to 2	2 to 3	25 (24,8 %)	61 (60,4 %)	15 (14,9 %)	<0.001	<0.001	0.86
	3 to 4	4 to 5	55 (54,5 %)	28 (27,7 %)	18 (17,8 %)	<0.001	0.005	0.66
* Based on regular two-sample z-tests								

As in experiment 1, there were significant differences between the two types of movements in all cases where a significant minority perceived no difference. For all three problems in group I, the symmetric comparisons, the results in experiment 2 were similar to the results in experiment 1. Hence, the perceived symmetry of the scale does not seem to be affected by the way the verbal anchors were used. In group II, a majority of subjects perceived differences. However, the differences are not of the same nature as in experiment 1. Significantly more subjects now perceive 3–4 as larger than 1–2, and 2–3 as larger than 4–5, in contrast to the previous situation when only the end points had verbal anchors. Similar changes can be seen in groups III and IV, where the movements between points near the middle of the scale are perceived as larger than movements near the end of the scale by significantly more subjects. Hence, the results in experiment 2 differ significantly from the results in experiment 1 since movements between points in the middle of the scale are experienced as greater changes in opinion than movements between points near an end of the scale in experiment 2. We will refer to this as the *middle-of-scale effect*. It should also be noted that the effect sizes were generally larger when all points on the scale had verbal anchors.

The main observation from these experiments is that there are systematic differences in the way respondents perceive the five-point Likert-type scale, depending on how verbal anchors are used. A possible explanation for why subjects perceive a scale with verbal anchors at the end points differently from a scale with verbal anchors at all points is that with the latter, it becomes more obvious that while a movement in the middle of the scale signals a change of opinion, a movement at the end of the scale only signals a change of intensity within the same opinion. With verbal anchors only at the ends, the subjects have to imagine the meaning of the other scale points themselves. Hence, without verbal anchors, it is not obvious that a movement between, for example, 2 and 4 is equivalent to a change in opinion.

There has been some previous research dealing with the effects of verbal anchors that these results can be related to (see Weijters et al., 2010, for a review). For example, verbal anchors are often assumed to make scale points more salient, which might attract respondents (Krosnick and Fabrigar, 1997). Hence, verbal anchors on intermediate points would create a shift towards those points at the expense of the extreme points (Simonson, 1989). The end-of-scale effect may also be seen as an explanation of the well-known central tendency bias (e.g. James, Demaree, and Wolf, 1984) in surveys. The relatively large perceived distance between points at the end of the scale makes it relatively hard to reach an end point when verbal anchors are used only at the end points. The perceived distance between points in the middle of the scale is smaller, which makes it easier to move between them.

The variance for a set of data of this type also depends on the way the verbal anchors are used, which should be taken into account in the statistical analysis. A rank-based test procedure (e.g. the Kruskal-Wallis test) is obviously immune to potential scale effects of the type we have discussed here; however, a parametric test procedure (e.g. the one-way ANOVA) is not. It is easy to see that assigning, for example, the values {0, 1.5, 2, 2.5, 4}, representing end-of-scale effect, to the five scale points will create a lower variance than, for example, the values {0, 0.5, 2, 3.5, 4}, representing middle-of-scale effect. Parametric test procedures are obviously affected by this problem.

Since neither of the two types of verbal anchor use creates a response scale that is perceived as equidistant, there are potential validity issues related to the choice of statistical methodology used in the analyses of data collected with Likert-type scales. In the next section, we will use a Monte Carlo approach to evaluate these issues.

3. Simulations

Design

An experimental design with three populations (k = 3) and four different combinations of small (defined as n = 5) and large (defined as n = 25) sample sizes were used. The simulations were based on random numbers from binomial distributions with a sample space {0, 1, 2, 3, 4} to represent five-level Likert-items, where $\mu_1 \leq \mu_2 \leq \mu_3$ and $|\mu_1 - \mu_2| = |\mu_3 - \mu_2|$. Symmetry was defined as a situation where the second population had a true mean value of 2. Moderate skewness was assumed to correspond with a true mean value of 1 for the second population, and severe skewness with a true mean value of 0.5. The three cases are illustrated in figure 3, with a symmetric distribution to the left, a moderately skewed distribution in the middle, and a severely skewed distribution to the right.

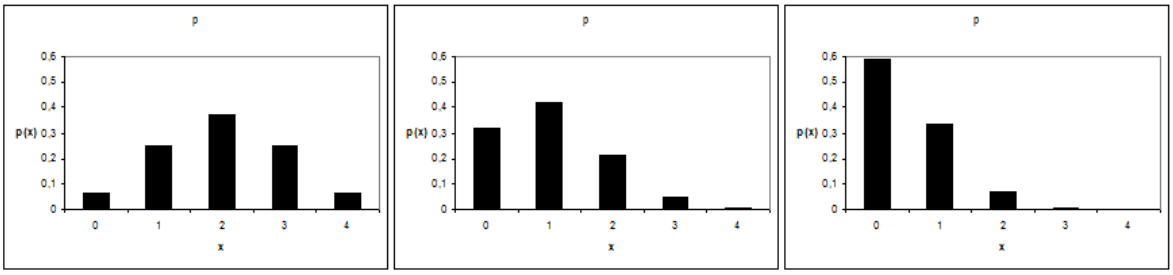

Figure 3: Symmetry, moderate skewness, and severe skewness

Table 3 shows the manner in which the true mean values of the distributions were shifted to achieve a suitable range of effect sizes (Cohen, 1992), ranging from no effect (f = 0) to a very large effect (f = 0.65). All mean values were calculated using G*Power version 3.1.2 (Faul et al., 2007). Note that the first population in the heavily skewed case with a very large effect size is technically not based on the binomial distribution as it consists exclusively of zeros.

Table 3: True mean values and implied effect sizes

Symmetrical			Moderately skewed			Heavily skewed			Effect size
μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	
2.000	2.000	2.000	1.000	1.000	1.000	0.500	0.500	0.500	f = 0.00
1.877	2.000	2.123	0.894	1.000	1.106	0.419	0.500	0.581	f = 0.10
1.696	2.000	2.304	0.737	1.000	1.263	0.299	0.500	0.701	f = 0.25
1.519	2.000	2.481	0.583	1.000	1.417	0.182	0.500	0.818	f = 0.40
1.243	2.000	2.757	0.344	1.000	1.656	0.000	0.500	1.000	f = 0.65

Four different test procedures were examined; the omnibus one-way ANOVA, the Kruskal-Wallis test (Kruskal and Wallis, 1952), the Brown-Forsythe test (Brown and Forsythe, 1974), and the Welch test (Welch, 1951). Every combination of test procedure and parent distribution was evaluated for every combination of sample sizes and effect size. For each combination, 3 × 50,000 sets of random numbers were generated, and the null hypothesis that corresponds to no difference between the locations of the populations was challenged at an alpha level of 0.05. The end-of-scale effect was simulated by adjusting the original scale to {0, 1.5, 2, 2.5, 4} and the middle-of-scale effect by adjusting it to {0, 0.5, 2, 3.5, 4}. All simulation procedures were conducted using Microsoft Excel 2010.

The null hypothesis of no difference in the true proportion of rejected tests between the four procedures for a certain combination was evaluated with chi-square tests in all cases. In cases where this null hypothesis was rejected, post hoc analysis was performed using pair-wise chi-square tests with Bonferroni correction.

Results

Table 4 shows the simulation results in the symmetric case with equidistant data. At (25, 25, 25), there is a significant difference between the methods only when the effect size is medium, where ANOVA (A) and the Brown-Forsythe test (BF) both are significantly more powerful than the Kruskal-Wallis test (KW). At (5, 5, 5), the methods differ significantly at all effect sizes. The overall picture from the detailed analyses is that A works best and that the Welch test (W) should be avoided in these circumstances. We also note that all methods except A are quite conservative and generate very few type I errors when all sample sizes are small. At (5, 5, 25), A is superior as well; however, it should be noted that KW dominates both BF and W at medium and larger effect sizes. At (5, 25, 25), KW has the lowest power when the effect size is small; however, when the effect size becomes larger, W performs worse while A and KW are more powerful. Hence, in the symmetric case with equidistant data, A should be recommended as it dominates the other three methods for many combinations of effect sizes and sample sizes, especially where at least one sample size is small, and is never dominated by any of them. This is obviously in line with what one could have expected according to theory, as data are metric and approximate homoscedasticity prevails in the symmetric case.

Table 5 shows the results from the symmetric case where the end-of-scale (EOS) effect is present. At (25, 25, 25), KW dominates W when the effect size is small. Furthermore, KW dominates all the other methods at medium and large effect sizes. The tendency is similar, but reinforced, when at least one sample size is small: KW dominates the other methods, and BF and W are weak. Moreover, when at least one sample size is small, BF and W are quite conservative, which is the probable reason for their low power. Hence, in the symmetric case when the EOS effect is present, KW should be the recommended method as it dominates the other three methods for most combinations of effect size and sample sizes. Note that the percentages for KW are the same as in the equidistant case above (and as in the middle-of-scale case below) as KW is not affected by non-equidistance since the ranks stay the same regardless of the severity of the equidistance assumption violation. Thus, the differences in percentages for the other methods illustrate the impact of the EOS effect when data are incorrectly assumed to be equidistant.

Table 4: Proportion of significant tests, symmetric data and equidistance

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
$n_1 = 25$	0.00	0.0503	0.0484	0.0499	0.0504	0.492
$n_2 = 25$	0.10	0.1105	0.1055	0.1099	0.1088	0.071
$n_3 = 25$	0.25	0.4605	0.4462	0.4598	0.4541	0.002
	0.40	0.8655	0.8564	0.8653	0.8602	0.343
	0.65	0.9992	0.9991	0.9992	0.9991	0.999
$n_1 = 5$	0.00	0.0488	0.0398	0.0428	0.0369	< 0.001
$n_2 = 5$	0.10	0.0584	0.0486	0.0514	0.0453	< 0.001
$n_3 = 5$	0.25	0.1088	0.0926	0.0964	0.0857	< 0.001
	0.40	0.2135	0.1838	0.1954	0.1725	< 0.001
	0.65	0.4878	0.4378	0.4584	0.4112	< 0.001
$n_1 = 5$	0.00	0.0516	0.0447	0.0492	0.0478	< 0.001
$n_2 = 5$	0.10	0.0701	0.0619	0.0650	0.0636	< 0.001
$n_3 = 25$	0.25	0.1927	0.1710	0.1518	0.1474	< 0.001
	0.40	0.4335	0.3890	0.3290	0.3162	< 0.001
	0.65	0.8455	0.8000	0.7009	0.6610	< 0.001
$n_1 = 5$	0.00	0.0513	0.0474	0.0557	0.0542	< 0.001
$n_2 = 25$	0.10	0.0751	0.0699	0.0789	0.0753	< 0.001
$n_3 = 25$	0.25	0.2297	0.2132	0.2154	0.1967	< 0.001
	0.40	0.5160	0.4915	0.4695	0.4398	< 0.001
	0.65	0.9100	0.8958	0.8574	0.8401	< 0.001

Table 5: Proportion of significant tests, symmetric data and end-of-scale effect

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
$n_1 = 25$	0.00	0.0499	0.0484	0.0493	0.0458	0.018
$n_2 = 25$	0.10	0.1015	0.1055	0.1003	0.0970	< 0.001
$n_3 = 25$	0.25	0.4163	0.4462	0.4140	0.4065	< 0.001
	0.40	0.8242	0.8564	0.8225	0.8167	< 0.001
	0.65	0.9973	0.9991	0.9973	0.9970	0.987
$n_1 = 5$	0.00	0.0336	0.0398	0.0206	0.0144	< 0.001
$n_2 = 5$	0.10	0.0418	0.0486	0.0257	0.0179	< 0.001
$n_3 = 5$	0.25	0.0789	0.0926	0.0516	0.0315	< 0.001
	0.40	0.1627	0.1838	0.1175	0.0656	< 0.001
	0.65	0.4085	0.4378	0.3343	0.1981	< 0.001
$n_1 = 5$	0.00	0.0546	0.0447	0.0262	0.0236	< 0.001
$n_2 = 5$	0.10	0.0704	0.0619	0.0395	0.0358	< 0.001
$n_3 = 25$	0.25	0.1627	0.1710	0.1212	0.0982	< 0.001
	0.40	0.3590	0.3890	0.2989	0.2496	< 0.001
	0.65	0.7620	0.8000	0.6637	0.6275	< 0.001
$n_1 = 5$	0.00	0.0543	0.0474	0.0420	0.0358	< 0.001
$n_2 = 25$	0.10	0.0726	0.0699	0.0580	0.0518	< 0.001
$n_3 = 25$	0.25	0.2065	0.2132	0.1709	0.1490	< 0.001
	0.40	0.4631	0.4915	0.4049	0.3640	< 0.001
	0.65	0.8656	0.8958	0.8207	0.7925	< 0.001

Table 6 shows the results from the symmetric case where the middle-of-scale (MOS) effect is present. At (25, 25, 25), there are no significant differences in performance between the methods; however, at (5, 5, 5) A dominates all the other methods at all effect sizes while KW generally performs worse than both BF and W. At (5, 5, 25), A continues to dominate the other methods at all effect sizes, while KW is now significantly more powerful than both BF and W at medium and larger effect sizes. At (5, 25, 25), BF and W are quite liberal and generate many type I errors, which also explains the superiority of BF and W at the small effect size. At larger effect sizes, BF and W, again, have less power than both A and KW, which in turn do not differ significantly. Hence, the MOS effect seems to affect BF and W more than it affects A. Therefore, in the symmetric case when the MOS effect is present, A should be recommended as it dominates the other three methods for most combinations of effect sizes and sample sizes.

Table 7 shows the results from the case where data are moderately skewed but equidistant. At (25, 25, 25), the only significant difference is that KW has less power at the medium effect size. At (5, 5, 5), all methods except A are excessively conservative, and A is more powerful while W is less powerful than KW and BF at all effect sizes. At (5, 5, 25) and (5, 25, 25), BF and W are more powerful than A and KW in most cases except for the very large effect size where W loses power, while KW is in turn generally more powerful than A. Hence, where data are moderately skewed but equidistant, A should be the recommended method under equal sample sizes as it is at least as powerful as any of the other methods in all situations. On the other hand, under unequal sample sizes, A should be avoided as it is the weakest of the methods in most cases. Here, BF must be recommended as it is the method with the highest power in most cases.

Table 6: Proportion of significant tests, symmetric data and middle-of-scale effect

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
n_1 = 25	0.00	0.0512	0.0484	0.0510	0.0511	0.156
n_2 = 25	0.10	0.1086	0.1055	0.1083	0.1082	0.391
n_3 = 25	0.25	0.4467	0.4462	0.4459	0.4435	0.878
	0.40	0.8529	0.8564	0.8525	0.8499	0.738
	0.65	0.9989	0.9991	0.9989	0.9987	0.999
n_1 = 5	0.00	0.0528	0.0398	0.0436	0.0458	< 0.001
n_2 = 5	0.10	0.0641	0.0486	0.0534	0.0554	< 0.001
n_3 = 5	0.25	0.1161	0.0926	0.0985	0.1034	< 0.001
	0.40	0.2225	0.1838	0.1918	0.2019	< 0.001
	0.65	0.4907	0.4378	0.4474	0.4549	< 0.001
n_1 = 5	0.00	0.0498	0.0447	0.0533	0.0610	< 0.001
n_2 = 5	0.10	0.0704	0.0619	0.0671	0.0785	< 0.001
n_3 = 25	0.25	0.1925	0.1710	0.1508	0.1625	< 0.001
	0.40	0.4327	0.3890	0.3167	0.3267	< 0.001
	0.65	0.8440	0.8000	0.6782	0.6457	< 0.001
n_1 = 5	0.00	0.0489	0.0474	0.0621	0.0623	< 0.001
n_2 = 25	0.10	0.0750	0.0699	0.0846	0.0849	< 0.001
n_3 = 25	0.25	0.2219	0.2132	0.2161	0.2059	< 0.001
	0.40	0.5023	0.4915	0.4590	0.4431	< 0.001
	0.65	0.8999	0.8958	0.8417	0.8289	< 0.001

Table 7: Proportion of significant tests, moderately skewed data and equidistance

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
n_1 = 25	0.00	0.0510	0.0486	0.0503	0.0496	0.364
n_2 = 25	0.10	0.1089	0.1046	0.1083	0.1071	0.168
n_3 = 25	0.25	0.4609	0.4453	0.4587	0.4589	< 0.001
	0.40	0.8714	0.8585	0.8702	0.8706	0.089
	0.65	0.9993	0.9990	0.9993	0.9992	0.999
n_1 = 5	0.00	0.0500	0.0423	0.0436	0.0317	< 0.001
n_2 = 5	0.10	0.0592	0.0513	0.0511	0.0393	< 0.001
n_3 = 5	0.25	0.1064	0.0929	0.0929	0.0714	< 0.001
	0.40	0.2100	0.1876	0.1838	0.1417	< 0.001

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
	0.65	0.4862	0.4629	0.4294	0.2950	< 0.001
n_1 = 5	0.00	0.0487	0.0431	0.0514	0.0458	< 0.001
n_2 = 5	0.10	0.0542	0.0591	0.0784	0.0767	< 0.001
n_3 = 25	0.25	0.1391	0.1516	0.1890	0.1847	< 0.001
	0.40	0.3397	0.3639	0.3925	0.3831	< 0.001
	0.65	0.7885	0.8055	0.7824	0.6599	< 0.001
n_1 = 5	0.00	0.0488	0.0456	0.0551	0.0520	< 0.001
n_2 = 25	0.10	0.0652	0.0672	0.0882	0.0844	< 0.001
n_3 = 25	0.25	0.1916	0.1952	0.2444	0.2230	< 0.001
	0.40	0.4523	0.4573	0.5231	0.4625	< 0.001
	0.65	0.8792	0.8818	0.9113	0.7315	< 0.001

Table 8 shows the results from the case where data are moderately skewed and the EOS effect is present. At (25, 25, 25), KW is significantly more powerful than any other method at medium and large effect sizes. At (5, 5, 5), all methods except A are too conservative, an effect that stays as the effect size becomes larger. W is less powerful than KW and BF at all effect sizes, and KW in turn dominates BF. At (5, 5, 25) and (5, 25, 25), W is exceedingly liberal while A is more powerful than the other methods in most cases except KW at large effect sizes. KW is also generally more powerful than BF. Hence, when all sample sizes are large, KW should be recommended if data are moderately skewed and the EOS effect is present. However, A should be the preferred method when at least one sample size is small as it then dominates the other three methods for most combinations of effect sizes and sample sizes.

Table 8: Proportion of significant tests, moderately skewed data and end-of-scale effect

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
n_1 = 25	0.00	0.0504	0.0486	0.0501	0.0514	0.282
n_2 = 25	0.10	0.1051	0.1046	0.1043	0.1038	0.936
n_3 = 25	0.25	0.4225	0.4453	0.4213	0.4174	< 0.001
	0.40	0.8365	0.8585	0.8357	0.8316	< 0.001
	0.65	0.9985	0.9990	0.9985	0.9984	0.999
n_1 = 5	0.00	0.0506	0.0423	0.0347	0.0300	< 0.001
n_2 = 5	0.10	0.0599	0.0513	0.0428	0.0370	< 0.001
n_3 = 5	0.25	0.1054	0.0929	0.0790	0.0681	< 0.001
	0.40	0.2089	0.1876	0.1678	0.1370	< 0.001
	0.65	0.4958	0.4629	0.4334	0.2947	< 0.001
n_1 = 5	0.00	0.0465	0.0431	0.0474	0.0971	< 0.001
n_2 = 5	0.10	0.0736	0.0591	0.0538	0.0684	< 0.001
n_3 = 25	0.25	0.1853	0.1516	0.1136	0.0886	< 0.001
	0.40	0.4018	0.3639	0.2721	0.1845	< 0.001
	0.65	0.8167	0.8055	0.6965	0.4426	< 0.001
n_1 = 5	0.00	0.0477	0.0456	0.0608	0.1121	< 0.001
n_2 = 25	0.10	0.0757	0.0672	0.0715	0.0914	< 0.001
n_3 = 25	0.25	0.2079	0.1952	0.1705	0.1439	< 0.001
	0.40	0.4596	0.4573	0.3917	0.3098	< 0.001
	0.65	0.8731	0.8818	0.8314	0.6245	< 0.001

Table 9 shows the results from the case where data are moderately skewed and the MOS effect is present. At (25, 25, 25), KW is significantly more powerful than any other method at medium and large effect sizes, exactly as under the EOS effect. At (5, 5, 5), BF and W are very conservative, and A and KW have more power than BF and W at all effect sizes. When the effect size is large or very large, KW also has more power than A. At (5, 5, 25) and (5, 25, 25), W becomes excessively liberal, which suggests that it should be avoided. It should also be noted that the MOS effect actually reduces the percentage of significant As at (5, 5, 25) when the effect size increases from none to a small size. However, BF generally has more power than both A and KW without resulting in a large number of type I errors. KW is in turn also more powerful than A. Hence, when data are moderately skewed and the MOS effect is present, KW should be the preferred choice when sample sizes are equal and BF when sample sizes are unequal.

Table 9: Proportion of significant tests, moderately skewed data and middle-of-scale effect

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
n_1 = 25	0.00	0.0498	0.0486	0.0486	0.0535	0.001
n_2 = 25	0.10	0.1019	0.1046	0.0998	0.1064	0.006
n_3 = 25	0.25	0.4197	0.4453	0.4164	0.4260	< 0.001
	0.40	0.8322	0.8585	0.8298	0.8357	< 0.001
	0.65	0.9982	0.9990	0.9981	0.9984	0.999
n_1 = 5	0.00	0.0435	0.0423	0.0202	0.0150	< 0.001
n_2 = 5	0.10	0.0491	0.0513	0.0240	0.0201	< 0.001
n_3 = 5	0.25	0.0893	0.0929	0.0469	0.0396	< 0.001
	0.40	0.1689	0.1876	0.0967	0.0840	< 0.001
	0.65	0.3809	0.4629	0.2437	0.1922	< 0.001
n_1 = 5	0.00	0.0463	0.0431	0.0585	0.1266	< 0.001
n_2 = 5	0.10	0.0313	0.0591	0.0961	0.2174	< 0.001
n_3 = 25	0.25	0.0684	0.1516	0.2160	0.4026	< 0.001
	0.40	0.1956	0.3639	0.4069	0.5755	< 0.001
	0.65	0.6178	0.8055	0.7332	0.7021	< 0.001
n_1 = 5	0.00	0.0466	0.0456	0.0546	0.1541	< 0.001
n_2 = 25	0.10	0.0503	0.0672	0.0915	0.2374	< 0.001
n_3 = 25	0.25	0.1433	0.1952	0.2575	0.4028	< 0.001
	0.40	0.3581	0.4573	0.5365	0.5794	< 0.001
	0.65	0.7878	0.8818	0.9089	0.7390	< 0.001

Table 10 shows the results from the case where data are heavily skewed but equidistant. At (25, 25, 25), W is somewhat more liberal than BF and more powerful at small or medium effect sizes. KW is less powerful than all other methods at medium effect size. At (5, 5, 5), A and KW are more powerful than BF and W at all effect sizes, and A is also more powerful than KW at a very large effect size. At (5, 5, 25) and (5, 25, 25), BF is predominant at all effect sizes and KW in turn dominates A. It should be noted that W is completely ineffective in case of a very large effect size, since all observed values in one of the three groups become equal to the end point of the scale, corresponding with a zero variance. W is also generally inefficient when at least one sample size is small, as the probability of a zero variance in at least one group increases. Hence, when data are heavily skewed but equidistant, A should be the recommended method at equal sample sizes, while BF should be preferred when sample sizes are unequal. In other words, the recommendation for equidistant data under heavy skewness is identical to the one under moderate skewness.

Table 11 shows the results from the case where data are heavily skewed and the EOS effect is present. At (25, 25, 25), there are no significant differences in performance between the methods; however, at (5, 5, 5), A dominates the other methods at all effect sizes while KW generally performs better than BF and W. At (5, 5, 25) and (5, 25, 25), BF is again generally predominant at all effect sizes; however, A dominates KW. Thus, when data are heavily skewed and the EOS effect is present, A should be the recommended method at equal sample sizes, while BF should be preferred when sample sizes are unequal. Note that these recommendations deviate from those under moderate skewness.

Table 10: Proportion of significant tests, heavily skewed data and equidistance

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
n_1 = 25	0.00	0.0481	0.0480	0.0473	0.0514	0.016
n_2 = 25	0.10	0.1082	0.1041	0.1065	0.1120	0.001
n_3 = 25	0.25	0.4569	0.4391	0.4529	0.4646	< 0.001
	0.40	0.8879	0.8757	0.8853	0.8906	0.066
	0.65	1.0000	1.0000	1.0000	0.0000	< 0.001
n_1 = 5	0.00	0.0435	0.0421	0.0365	0.0077	< 0.001
n_2 = 5	0.10	0.0529	0.0510	0.0438	0.0095	< 0.001
n_3 = 5	0.25	0.0930	0.0900	0.0765	0.0165	< 0.001
	0.40	0.1839	0.1839	0.1511	0.0269	< 0.001
	0.65	0.4487	0.4813	0.3633	0.0000	< 0.001

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
n_1 = 5	0.00	0.0468	0.0417	0.0479	0.0086	< 0.001
n_2 = 5	0.10	0.0365	0.0414	0.0852	0.0082	< 0.001
n_3 = 25	0.25	0.0814	0.1092	0.2126	0.0278	< 0.001
	0.40	0.2219	0.2895	0.4347	0.0838	< 0.001
	0.65	0.6906	0.8133	0.8268	0.0000	< 0.001
n_1 = 5	0.00	0.0471	0.0441	0.0582	0.0163	< 0.001
n_2 = 25	0.10	0.0526	0.0557	0.1017	0.0190	< 0.001
n_3 = 25	0.25	0.1518	0.1605	0.2682	0.0632	< 0.001
	0.40	0.3735	0.3941	0.5697	0.1501	< 0.001
	0.65	0.8301	0.8769	0.9824	0.0000	< 0.001

Table 11: Proportion of significant tests, heavily skewed data and end-of-scale effect

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
n_1 = 25	0.00	0.0500	0.0480	0.0496	0.0520	0.046
n_2 = 25	0.10	0.1074	0.1041	0.1066	0.1083	0.196
n_3 = 25	0.25	0.4451	0.4391	0.4437	0.4480	0.206
	0.40	0.8769	0.8757	0.8760	0.8758	0.997
	0.65	1.0000	1.0000	1.0000	0.0000	< 0.001
n_1 = 5	0.00	0.0532	0.0421	0.0385	0.0102	< 0.001
n_2 = 5	0.10	0.0640	0.0510	0.0469	0.0126	< 0.001
n_3 = 5	0.25	0.1114	0.0900	0.0844	0.0215	< 0.001
	0.40	0.2198	0.1839	0.1758	0.0330	< 0.001
	0.65	0.5563	0.4813	0.4602	0.0000	< 0.001
n_1 = 5	0.00	0.0468	0.0417	0.0445	0.0144	< 0.001
n_2 = 5	0.10	0.0510	0.0414	0.0654	0.0107	< 0.001
n_3 = 25	0.25	0.1321	0.1092	0.1575	0.0124	< 0.001
	0.40	0.3346	0.2895	0.3569	0.0275	< 0.001
	0.65	0.8603	0.8133	0.8761	0.0000	< 0.001
n_1 = 5	0.00	0.0470	0.0441	0.0678	0.0186	< 0.001
n_2 = 25	0.10	0.0607	0.0557	0.1079	0.0179	< 0.001
n_3 = 25	0.25	0.1700	0.1605	0.2597	0.0477	< 0.001
	0.40	0.4087	0.3941	0.5361	0.1060	< 0.001
	0.65	0.8946	0.8769	0.9946	0.0000	< 0.001

Finally, table 12 shows the results from the case where data are heavily skewed and the MOS effect is present. At (25, 25, 25), KW dominates all other methods at all effect sizes, except when the effect size is small, in which case it does not differ significantly from W. At (5, 5, 5), KW is also predominant at all effect sizes. At (5, 5, 25) and (5, 25, 25), however, BF has higher power than the other methods at all effect sizes except the very large one. Hence, under heavy skewness and MOS effect, A should be the recommended method at equal sample sizes, while BF should be preferred when sample sizes are unequal. Again, these recommendations are identical to the recommendations under moderate skewness.

Table 12: Proportion of significant tests, heavily skewed data and middle-of-scale effect

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
n_1 = 25	0.00	0.0434	0.0480	0.0400	0.0553	< 0.001
n_2 = 25	0.10	0.0898	0.1041	0.0848	0.1080	< 0.001
n_3 = 25	0.25	0.3764	0.4391	0.3658	0.4126	< 0.001
	0.40	0.8059	0.8757	0.7976	0.8369	< 0.001
	0.65	0.9999	1.0000	0.9998	0.0000	< 0.001
n_1 = 5	0.00	0.0292	0.0421	0.0143	0.0007	< 0.001
n_2 = 5	0.10	0.0355	0.0510	0.0165	0.0008	< 0.001
n_3 = 5	0.25	0.0602	0.0900	0.0282	0.0026	< 0.001
	0.40	0.1151	0.1839	0.0500	0.0049	< 0.001
	0.65	0.2746	0.4813	0.1167	0.0000	< 0.001

Sample sizes	Effect size	Test procedure				p-value
		ANOVA	Kruskal-Wallis	Brown-Forsythe	Welch	
$n_1 = 5$	0.00	0.0576	0.0417	0.0668	0.0041	< 0.001
$n_2 = 5$	0.10	0.0279	0.0414	0.1453	0.0162	< 0.001
$n_3 = 25$	0.25	0.0173	0.1092	0.3330	0.0762	< 0.001
	0.40	0.0405	0.2895	0.5322	0.1754	< 0.001
	0.65	0.2082	0.8133	0.7263	0.0000	< 0.001
$n_1 = 5$	0.00	0.0479	0.0441	0.0462	0.0183	< 0.001
$n_2 = 25$	0.10	0.0415	0.0557	0.0822	0.0417	< 0.001
$n_3 = 25$	0.25	0.1037	0.1605	0.2208	0.1343	< 0.001
	0.40	0.2536	0.3941	0.4668	0.2561	< 0.001
	0.65	0.6133	0.8769	0.8555	0.0000	< 0.001

4. Conclusion

The methodological issue of whether a Likert-type scale can be reasonably assumed to have metric properties and the type of statistical method that should consequently be used to analyse Likert-type data goes back a long time, and has been discussed in relation to a number of paradigms. Still, parametric methods are often used in contemporary research to analyse data that are not equidistant by nature (Jakobsson, 2004). In this study, we saw that respondents generally did not perceive a Likert-type scale as equidistant, and that the nature of the perceived non-equidistance depended on how verbal anchors were connected to the scale points. We also tested the sensitivity of common statistical methods to the two main types of non-equidistance under different circumstances. The overall conclusion from these simulations was that the best statistical method to compare different groups of Likert-type data seems to depend both on the expected scale effect, that is, the nature of the non-equidistance created by the use of verbal anchors and on the degree of skewness (see figure 4). Hence, this study contributes to the methodological literature in two different ways.

Skewness	Expected scale effect		
	Equidistance	End-of-scale	Middle-of-scale
Approximate symmetry	ANOVA	Kruskal-Wallis	ANOVA
Moderate skewness	ANOVA (if sample sizes are approximately equal) or Brown-Forsythe (otherwise)	Kruskal-Wallis (if all sample sizes are large) or ANOVA (otherwise)	Kruskal-Wallis (if sample sizes are approximately equal) or Brown-Forsythe (otherwise)
Heavy skewness		ANOVA (if sample sizes are approximately equal) or Brown-Forsythe (otherwise)	

Figure 4: The preferred method with respect to expected scale effect and skewness

Further research regarding subject perceptions of the Likert-type scale is required. The standard verbal anchors for the second and fourth scale points are 'Disagree' and 'Agree', respectively. Changing them to, for example, 'Somewhat disagree' and 'Somewhat agree' might change the way subjects perceive the scale in terms of distance between the points. Scales with more, or less, than five scale points should also be evaluated in a similar manner.

Further research is also required to evaluate the robustness of parametric methods to violations of the equidistance assumption. Other types of non-equidistant data can be used to evaluate the robustness of the parametric methods in more detail, as well as other types of statistical tests commonly used to analyse Likert-type data.

References

Albaum, G., Best, R. & Hawkins, D. (1977). The measurement properties of semantic scale data. *Journal of the Market Research Society*, 19 (1), 21-26.

Albaum, G. (1997). The Likert scale revisited: An alternate version. *Journal of the Market Research Society*, 39 (2), 331-348.

Alexandrov, A. (2010). "Characteristics of Single-Item Measures in Likert Scale Format". *The Electronic Journal of Business Research Methods*, 8 (1), 1-12.

Bendixen, M. T. & Sandler, M. (1995). Converting verbal scales to interval scales using correspondence analysis. *Management Dynamics: Contemporary Research*, 4 (1), 31-49.

Brown, M. B. & Forsythe, A. B. (1974). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics*, 30, 719-724.

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, 106-116.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.

Cunningham, W. H., Anderson Jr, W. T. & Murphy, J. H. (1974). Are students real people? *The Journal of Business*, 47 (3), 399-409.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.

Feir, B. J. & Toothaker, L. E. (1974). The ANOVA F-Test Versus The Kruskal-Wallis Test: A Robustness Study. *The Annual Meeting of the American Educational Research Association*, Chicago, Il.

Glass, G. V., Peckham, P. D. & Sanders, J. R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research*, 42 (3), 237–288.

Granberg-Rademacker, J. S. (2010). An Algorithm for Converting Ordinal Scale Measurement Data to Interval/Ratio Scale. *Educational and Psychological Measurement*, 70 (1), 74-90.

Harwell, M. R. & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, 71, 105-131.

Jakobsson, U. (2004). Statistical presentation and analysis of ordinal data in nursing research. *Scandinavian Journal of Caring Sciences*, 18, 437-440.

James, L. R., Demaree, R. G. & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69, 85-98.

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education*, 38, 1212-1218.

Kennedy, R., Riquier, C. & Sharp, B. (1996). Practical Applications of Correspondence Analysis to Categorical Data in Market Research. *Journal of Targeting, Measurement and Analysis for Marketing*, 5 (1), 56-70.

King, G., Murray, C. J. L., Salomon, J. A. & Tandon, A. (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review*, 98 (1), 191-207.

Krosnick, J. A. & Fabrigar, L. R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In: Lyberg, L.E., Biemer, P., Collins, M., de Leeuw, E.D., Dippo, C., Schwarz, N., and Trewin, D. (Eds.). *Survey Measurement and Process Quality*. NY: Wiley-Interscience.

Kruskal, W. H. & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis, *Journal of the American Statistical Association*, 47 (260), 583–621.

Lantz, B. (2012). The impact of sample non-normality on ANOVA and alternative methods, forthcoming in *British Journal of Mathematical and Statistical Psychology*. DOI: 10.1111/j.2044-8317.2012.02047.x

Lee, J. A. & Soutar, G. N. (2010), Is Schwartz's value survey an interval scale, and does it really matter?" *Journal of Cross-Cultural Psychology*, 41, 76-86.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology* 140, 1-55.

Michell, J. (1986). Measurement Scales and Statistics: A Clash of Paradigms. *Psychological Bulletin*, 100 (3), 398-407.

Mundy J. & Dickinson D. (2004). Factors affecting the uptake of voluntary HIV/AIDS counselling and testing (VCT) services in the workplace. *In: HIV/AIDS in the Workplace Research Symposium*. University of Witwatersrand, June 2004, 175-193.

Pearse, N. (2011). Deciding on the Scale Granularity of Response Categories of Likert type

Scales: The Case of a 21-Point Scale. *The Electronic Journal of Business Research Methods*, 9 (2), 159-171.

Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effect. *Journal of Consumer Research*, 16, 158-174.

Stevens, S. S. (1946). On the Theory of Scales of Measurement, *Science,* 7 (103), 677-680.

Weijters, B., Cabooter, E. & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27, 236-247.

Welch, B. L. (1951). On the Comparison of Several Mean Values: An Alternative Approach. *Biometrika*, 38, 330–336.

Wu, C-H. (2007). An Empirical Study on the Transformation of Likert-scale Data to Numerical Scores. *Applied Mathematical Sciences*, 1, 58, 2851-2862.

Zimmerman, D. W. (1998). Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions, *Journal of Experimental Education*, 67:1, 55–68.