Methods for comparing the responses from a Likert question, with paired observations and independent observations in each of two samples

Ben Derrick, Paul White

Faculty of Environment and Technology, University of the West of England, Bristol, BS16 1QY, UK

Email: ben.derrick@uwe.ac.uk; paul.white@uwe.ac.uk

ABSTRACT

Researchers often encounter two samples of Likert data, which contain both independent observations and paired observations. Standard analyses in this scenario typically involve discarding the independent observations and performing the paired samples t-test, the Wilcoxon signed-rank test or the Pratt test. These naive approaches are examined alongside recently developed partially overlapping samples t-tests that make use of all of the available data in the two sample scenario. For two samples of observations from a Likert question with five categories or seven categories, test statistics are assessed for their Type I error robustness and power. A summary measure of Type I error robustness across the simulation design is quantified as that value of π such that (1-π)$\times$100 percent of Type I error rates are within $\pi \times$100 percent of the nominal significance level. Across a range of sample sizes and correlation coefficients, the partially overlapping samples t-tests are Type I error robust, and offer a more powerful alternative for the analysis of two samples including both paired observations and independent observations. In these scenarios, when the assumption of an underlying continuous distribution is not inappropriate, the partially overlapping samples t-test is recommended.

Keywords: Likert item; ordinal; partially overlapping samples; simulation; Type I error robustness

Mathematics Subject Classification: 60 62

1. INTRODUCTION

Situations arise when both paired observations and independent observations are present in a sample. This is referred to as a partially overlapping sample (Derrick *et al.*, 2015; Derrick, Toher and White, 2017). This paper evaluates tests used in the comparison of two partially overlapping samples. Previous literature in this area has focused on normally distributed data. The focus of this paper is for responses from an ordinal scale assuming equal spacing between the categories. The ordinal scales are represented by a five point Likert question, and a seven point Likert style question.

An example of two partially overlapping samples is a comparison of the responses of two Likert questions in the same survey, where some participants did not complete both questions, as obtained by Maisel and Fingerhut (2011). A further example of two partially overlapping samples for an individual Likert question is a comparison of the responses between pre-test and post-test, where some participants were not available at both times, as obtained by Bradley, Waliczek, and Zajicek

(1999). In both of these examples the authors discarded the unpaired observations and performed the paired samples *t*-test. Assuming data are missing completely at random (MCAR) this approach is not unjustified, given the large sample sizes obtained. However, power may be adversely affected for studies with smaller sample sizes.

Due to their intuitive appeal and simple construction, Likert questions are popular when measuring attitudes of respondents (Nunally, 1978). In certain methodological and practical aspects, Likert question responses may approximate interval level data, and can be analysed assuming an underlying continuous scale (Norman, 2010). Historically a Likert question consists of five options (Likert, 1932). The ordinal codes -2, -1, 0, 1, 2 could be applied to these options, with "0" representing the neutral response. In addition, seven point Likert style questions are commonly used, with ordinal codes -3, -2, -1, 0, 1, 2, 3.

Balanced and equally spaced response options around a neutral option are assumed for a valid Likert question (Uebersax, 2006). The exact wording of the neutral response is not an issue (Armstrong, 1987). If the options either side of the neutral response are not perceived to be balanced, then the assumption that responses approximate interval level data may not be reasonable (Bishop and Herron, 2015). Other issues with Likert questions include the responder tendency to give positive responses, and the potential for differing interpretation of the categorical options by both the responder and the analyst (Hodge and Gillespie, 2003). However, when the assumption of an underlying continuous distribution is not inappropriate and the questions are suitably formed, parametric tests for differences between the two sample means may be reasonable (Jamieson, 2004; Allen and Seaman, 2007; Derrick and White, 2017).

When comparing paired samples of ordinal data, the Wilcoxon signed-rank test can give dissimilar results to the paired samples *t*-test, and the correct choice of analysis depends on the exact form of the question of interest (Roberson *et al.*, 1994). Non-parametric tests are not inappropriate when interval approximating data is assumed, if the only potential difference between the samples is their central location (Clason and Dormody, 1994; Sisson and Stocker, 1989). Given the discrete nature of Likert questions, zero differences between pairs occur frequently. The Pratt (1959) test, which incorporates zero-differences in its calculation, can overcome the issues of the Wilcoxon signed-rank test which discards zero differences (Conover, 1973; Derrick and White, 2017).

The Pratt test, the Wilcoxon signed-rank test and the paired samples *t*-test are easily extended for the use when two partially overlapping samples are present, if the researcher is willing to discard any unpaired data. However, the discarding of data may introduce bias and reduce power and as is therefore a naïve approach (Guo and Yuan, 2017). Researchers should ensure that the analyses correctly reflect the design of the study. An alternative approach is the partially overlapping samples *t*-tests proposed by Derrick *et al.* (2017). These solutions are generalized forms of the *t*-test and have the advantage of making use of all of the available data. These solutions are Type I error robust under the assumptions of normality and MCAR, and are more powerful than alternatives that discard data (Derrick, Toher and White, 2017). The partially overlapping samples *t*-tests were previously considered for normally distributed data, the properties for ordinal data were not discussed.

The partially overlapping samples *t*-tests are given (Section 2) and demonstrated by example (Section 3). Methodology for comparing these proposals, the paired samples *t*-test, the Wilcoxon signed-rank test, and the Pratt test, is outlined for a five point Likert question and a seven point Likert style question (Section 4). Type I error robustness and power of the test statistics are assessed for scenarios where there are two partially overlapping samples (Section 5).

2. THE PARTIALLY OVERLAPPING SAMPLES *T*-TESTS

For situations comprising of a combination of both paired observations and unpaired observations for two samples, let 'n_a' represent the number of observations only in Group 1, and 'n_b' represent the number of observations only in Group 2 and 'n_c' represent the number of paired observations. It follows that the total sample size in Group 1 is $n_1 = n_a + n_c$, and the total sample size in Group 2 is $n_2 = n_b + n_c$.

There are two versions of the partially overlapping samples *t*-test, T_{new1} assumes equal variances between the two groups, and T_{new2} makes use of separate variances. For equal variances assumed the partially overlapping samples *t*-test by Derrick *et al.* (2017) is defined as:

$$T_{\text{new1}} = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2r\left(\frac{n_c}{n_1 n_2}\right)}}$$

where $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}}$ and *r* is Pearson's correlation coefficient.

The test statistic T_{new1} is referenced against the t-distribution with degrees of freedom:

$$v_1 = (n_c - 1) + \left(\frac{n_a + n_b + n_c - 1}{n_a + n_b + 2n_c}\right)(n_a + n_b).$$

For the comparison of samples on a continuous scale, when variances are unequal, Welch's test has superior Type I error robustness properties (Derrick, Toher and White, 2016). The statistic T_{new2} uses Welch's approximation to degrees of freedom and is defined by Derrick *et al.* (2017) as:

$$T_{\text{new2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} - 2r\left(\frac{S_1 S_2 n_c}{n_1 n_2}\right)}}$$

The test statistic T_{new2} is referenced against the *t*-distribution with degrees of freedom:

$$v_2 = (n_c - 1) + \left(\frac{\gamma - n_c + 1}{n_a + n_b + 2n_c}\right)(n_a + n_b) \text{ where } \gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 / (n_1 - 1) + \left(\frac{S_2^2}{n_2}\right)^2 / (n_2 - 1)}.$$

3. WORKED EXAMPLE

For a university undergraduate course, student satisfaction for a Mathematics module (Group 1) is compared against that of a Statistics module (Group 2). Most students under consideration study both modules, however some study only one. For each group, an online module evaluation is given to the students with the question "I am overall satisfied with the module". The answers obtained are given in Table 1.

Table 1. Responses to the question "I am overall satisfied with the module". Results coded as; Strongly Agree = 2, Agree = 1, Neither agree nor disagree = 0, Disagree = -1, Strongly Disagree = -2.

Unique ID	1	2	3	4	5	6	7	8	9	10	11	12	13
Group 1	0	0	2	0	-1		1	-1		0		1	2
Group 2	0		2	0	1	2	1	2	2	2	1		2

Using the convention of lower case for calculated sample values; $\bar{x}_1 - \bar{x}_2 =$-0.964, $n_a =$2, $n_b =$3, $n_c =$8, $T_{\text{new1}} =$-2.666, $v_1 =$9.857, $T_{\text{new2}} =$-2.609, $v_2 =$9.304. Using critical *t*-values at the 5% significance level (two sided) gives evidence to suggest than the module means are different. Students appear to be more satisfied with the Statistics module relative to the Mathematics module. Or more simply, using the R 'partiallyoverlapping' package (Derrick, 2017; Derrick, Toher and White, 2017), the p-values for T_{new1} and T_{new2} are 0.024 and 0.028 respectively.

Both forms of the partially overlapping samples t-test provide evidence to suggest that there is a significant difference between the two groups. This is in contrast to the conclusions that are made if performing standard tests that ignore the unpaired observations [paired samples *t*-test, *p*=0.088; Wilcoxon signed-rank test, *p*=0.174; Pratt test *p*=0.085].

4. METHODOLOGY

Monte-Carlo simulation methods are used to compare test statistics for the comparison of two samples which include both paired observations and unpaired observations. The standard tests considered are the paired samples *t*-test T_{paired}, the Wilcoxon signed-rank test W_1, and the Pratt test W_2. For each of these standard tests, randomly generated independent observations are ignored. The standard tests are compared against the proposals by Derrick *et al.* (2017), T_{new1} and T_{new2}. Small sample sizes are of particular interest.

The simulation is undertaken by discretizing realizations from bivariate Normal distributions to a five point scale and a seven point scale. This is done over a range of sample sizes {n_a, n_b, n_c} and non-negative Pearson's correlation coefficients {ρ}.

For the n_a independent observations in Group 1, the Mersenne-Twister algorithm (Matsumoto and Nishimura, 1998) generates pairs of random *U*(0,1) deviates and are transformed into Standard Normal deviates using the Box and Muller (1958) transformation. This process is repeated to generate the n_b independent observations in Group 2. For the n_c paired observations, additional Standard Normal deviates are generated, and these are transformed into correlated Standard Normal bivariates using methodology outlined by Kenney and Keeping (1951).

Let Standard Normal deviates be y_{ij} to denote the *i*-th observation in group *j*. Without loss of generality, for a five point scale the points are numbered from -2 to 2. The responses x_{ij} are calculated using the cut-points as follows:

$$x_{ij} = \begin{cases} 2 & \text{if} \quad y_{ij} > 0.8416 \\ 1 & \text{if} \quad 0.2533 \le y_{ij} \le 0.8416 \\ 0 & \text{if} \quad -0.2533 \le y_{ij} \le 0.2533 \\ -1 & \text{if} \quad -0.8416 \le y_{ij} \le -0.2533 \\ -2 & \text{if} \quad y_{ij} < -0.8416 \end{cases}$$

The cut-points are calculated so that under the Standard Normal distribution the theoretical distribution of the responses is uniform. Similarly, for a seven point scale, x_{ij} are calculated using the cut-points as follows:

$$x_{ij} = \begin{cases} 3 & \text{if} \quad y_{ij} > 1.6757 \\ 2 & \text{if} \quad 0.5659 \le y_{ij} \le 1.6757 \\ 1 & \text{if} \quad 0.1800 \le y_{ij} \le 0.5659 \\ 0 & \text{if} \quad -0.1800 \le y_{ij} \le 0.1800 \\ -1 & \text{if} \quad -0.5659 \le y_{ij} \le -0.1800 \\ -2 & \text{if} \quad -1.6757 \le y_{ij} \le -0.5659 \\ -3 & \text{if} \quad y_{ij} < -1.6757 \end{cases}$$

The median of Group 1 and the median of Group 2 are represented by η_1 and η_2 respectively. The scenarios compared encompass each integer value of η_1 and η_2. For example, by symmetry the

Type I error robustness when $\eta_1 = \eta_2 = 1$ is equivalent to Type I error robustness where $\eta_1 = \eta_2 =$ -1. The complete list of scenarios, parameters and the test statistics compared, can be found in Table 2. The scenarios and parameter combination are considered as part of a factorial design. For each scenario and parameter combination, the number generating process is repeated 10,000 times. For each repetition the null hypothesis is assessed at the $\alpha = 5\%$ significance level, two sided.

Table 2. Simulation design

Sample size	$n_a =$ (5, 10, 20, 30), $n_b =$ (5, 10, 20, 30), $n_c =$ (5, 10, 20, 30)									
ρ	0.00, 0.25, 0.50, 0.75									
Scenarios under $\eta_1 = \eta_2$	Five point scale					Seven point scale				
		μ_1	μ_2	η_1	η_2		μ_1	μ_2	η_1	η_2
	i)	0	0	0	0	*x*)	0	0	0	0
	ii)	0.5244	0.5244	1	1	*xi*)	0.3661	0.3661	1	1
	iii)	1.2816	1.2816	2	2	*xii*)	0.7916	0.7916	2	2
						xiii)	1.4652	1.4652	3	3
Scenarios under $\eta_1 \neq \eta_2$	Five point scale					Seven point scale				
		μ_1	μ_2	η_1	η_2		μ_1	μ_2	η_1	η_2
	iv)	0	0.5244	0	1	*xiv*)	0	0.3661	0	1
	v)	0	1.2816	0	2	*xv*)	0	0.7916	0	2
	vi)	0.5244	1.2816	1	2	*xvi*)	0	1.4652	0	3
	vii)	-0.5244	0.5244	-1	1	*xvii*)	0.3661	0.7916	1	2
	viii)	-0.5244	1.2816	-1	2	*xviii*)	0.3661	1.4652	1	3
	ix)	-1.2816	1.2816	-2	2	*xix*)	0.7916	1.4652	2	3
						xx)	-0.3661	0.3661	-1	1
						xxi)	-0.3661	0.7916	-1	2
						xxii)	-0.3661	1.4652	-1	3
						xxiii)	-0.7916	0.7916	-2	2
						xxiv)	-0.7916	1.4652	-2	3
						xxv)	-1.4652	1.4652	-3	3
Test Statistics	T_{paired}	Paired samples *t*-test								
	W_1	Wilcoxon signed-rank test (Standard method, discarding zeroes)								
	W_2	Pratt test (Wilcoxon signed-rank test, with Pratt's zeroes modification)								
	T_{new1}	Partially overlapping samples *t*-test with equal variances								
	T_{new2}	Partially overlapping samples *t*-test with unequal variances.								

Number of iterations: 10,000
Significance level: $\alpha = 0.05$

All calculations are performed in R. The paired samples *t*-test is calculated using the 'stats' package (R core team, 2015). The Wilcoxon signed-rank test is calculated using the Normal approximation corrected for ties with continuity correction factor, using the 'stats' package (R core team, 2015). The Pratt test is calculated under the same conditions using the 'coin' package (Hothorn, 2017). The partially overlapping samples *t*-tests are calculated using the 'partiallyoverlapping' package (Derrick, 2017).

5. RESULTS

For each parameter combination where $\eta_1 = \eta_2$, the proportion of the 10,000 iterations where the null hypothesis is rejected, represents the Type I error rate of the test under those conditions.

For selected parameter combinations, Type I error rates are given in Table 3. Liberal robustness criteria by Bradley (1978), offers guidance for assessing the Type I error rate for a given parameter combination. Under this criteria, Type I error robust statistics are within 50% of the nominal Type I error rate. For each parameter combination given in the table, the Type I error rates where 0.025 $\leq \alpha \leq$ 0.075 are highlighted in bold.

A summary measure of Type I error robustness across the entire simulation design for each of the test statistics is additionally put forward. The overall Type I error robustness is quantified as that value of π such that (1-π)×100 percent of Type I error rates are within π×100 percent of α. Large values of (1-π) are desirable. Table 3 shows the overall Type I error robustness of each of the test statistics.

Table 3. Type I error rates for selected parameter combinations, and overall robustness (1-π) across the simulation design, where $\eta_1 = \eta_2$.

	η_1	η_2	n_a	n_b	n_c	ρ	T_{paired}	W_1	W_2	T_{new1}	T_{new2}
Five point scale	0	0	5	5	5	0.5	**.042**	.010	.019	**.041**	**.038**
	0	0	5	20	10	0.5	**.052**	**.033**	**.051**	**.041**	**.044**
	1	1	5	5	5	0.5	**.040**	.006	.013	**.048**	**.041**
	1	1	5	20	10	0.5	**.045**	**.028**	**.049**	**.046**	**.050**
	2	2	5	5	5	0.5	.010	.001	.003	**.037**	.024
	2	2	5	20	10	0.5	**.027**	.001	**.044**	**.041**	**.052**
Value of (1-π) over all parameter combinations							.747	.584	.721	.821	.814
Seven point scale	0	0	5	5	5	0.5	**.046**	.008	.023	**.042**	**.048**
	0	0	5	20	10	0.5	**.049**	**.034**	**.048**	**.050**	**.050**
	1	1	5	5	5	0.5	**.044**	.005	.020	**.046**	**.041**
	1	1	5	20	10	0.5	**.049**	**.035**	**.051**	**.047**	**.047**
	2	2	5	5	5	0.5	**.026**	.001	.008	**.426**	**.032**
	2	2	5	20	10	0.5	**.041**	**.027**	**.046**	**.042**	**.045**
	3	3	5	5	5	0.5	.006	.000	.002	**.041**	.022
	3	3	5	20	10	0.5	**.026**	.001	**.055**	**.048**	**.053**
Value of (1-π) over all parameter combinations							.793	.622	.749	.871	.848

Table 3 shows that T_{new1} performs within Bradley's liberal Type I error robustness criteria, this remains true for the smallest sample size combination within the simulation design. For each of the other test statistics considered, Type I error robustness is not always maintained when both groups are heavily skewed. The Pratt test better controls for Type I error rates than the standard Wilcoxon signed-rank test.

In summary, for a five point Likert scale, the paired samples *t*-test is 74.7% robust. This means that 74.7% of the Type I error rates for parameter combinations within the simulation design are within

25.3% of the nominal Type I error rate. In this design, the paired samples *t*-test therefore maintains greater Type I error robustness than the Wilcoxon test (62.3%) or the Pratt test (70.6%). Both T_{new1} and T_{new2} are over 80% robust, thus maintain Type I error robustness better than the other tests considered.

The Type I error rates follow a similar pattern whether a five point scale or a seven point scale is used.

For all parameter combinations where $\eta_1 \neq \eta_2$, the percentage of iterations where the null hypothesis is rejected, represents the power of the test. Table 4 summarizes the power for each scenario using each test statistic, averaged over all parameter combinations.

Table 4. Power for each test statistic averaged over all scenarios where $\eta_1 \neq \eta_2$.

	Scenario	η_1	η_2	T_{paired}	W_1	W_2	T_{new1}	T_{new2}
Five point scale	*iv*	0	1	.380	.327	.358	.530	.524
	v	0	2	.788	.679	.740	.972	.966
	vi	1	2	.503	.440	.492	.734	.723
	vii	-1	1	.744	.642	.698	.937	.931
	viii	-1	2	.916	.746	.855	.998	.998
	ix	2	-2	.983	.779	.946	1.000	1.000
	Average			**.747**	**.623**	**.706**	**.882**	**.877**
Seven point scale	*xiv*	0	1	.244	.206	.227	.323	.319
	xv	0	2	.611	.536	.579	.821	.812
	xvi	0	3	.840	.713	.794	.989	.986
	xvii	1	2	.285	.244	.269	.392	.387
	xviii	1	3	.713	.628	.682	.933	.923
	xix	2	3	.433	.384	.431	.646	.634
	xx	-1	1	.582	.509	.550	.782	.774
	xxi	-1	2	.794	.677	.752	.964	.959
	xxii	-1	3	.918	.743	.871	.999	.998
	xxiii	-2	2	.899	.733	.856	.996	.995
	xxiv	-2	3	.967	.754	.931	1.000	1.000
	xxv	-3	3	.994	.773	.977	1.000	1.000
	Average			**.690**	**.575**	**.660**	**.820**	**.816**

Table 4 shows that T_{new1} and T_{new2} both consistently out-perform the standard tests which discard data. T_{new1} demonstrates marginally superior Type I error robustness and power properties relative to T_{new2}.

6. CONCLUSION

This paper has used simulation to compare the performance of test statistics where there are two samples, each sample with both paired observations and independent observations. This comparison has been performed for ordinal data, specifically for responses from either a five point Likert question, or a seven point Likert style question. Assuming the responses represent interval data, standard approaches such as the paired samples *t*-test or the Pratt test may not be inappropriate. However, these standard approaches discard the independent observations and as such are less than ideal, particularly if the sample sizes are small.

The partially overlapping samples *t*-tests proposed by Derrick *et al.* (2017) overcome the issue of discarding data. It is demonstrated that T_{new1} exhibits superior Type I error robustness relative to the other test statistics considered, and also has greater power. Therefore when the underlying assumptions of interval data are met, T_{new1} is recommended as the test of choice when comparing the responses from a Likert question with paired observations and independent observations in each of two samples.

7. REFERENCES

Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality Progress*, **40(7)**, 64-65.

Armstrong, R. L. (1987). The midpoint on a five-point Likert-type scale. *Perceptual and Motor Skills*, **64(2)**, 359-362. doi.org/10.2466/pms.1987.64.2.359

Bishop, P. A., & Herron, R. L. (2015). Use and misuse of the Likert item responses and other ordinal measures. *International Journal of Exercise Science*, **8(3)**, 297-302.

Box, G. E. P., & Muller, M. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, **29**, 610-611.

Bradley, J. C., Waliczek, T. M., & Zajicek, J. M. (1999). Relationship between environmental knowledge and environmental attitude of high school students. *The Journal of Environmental Education*, **30(3)**, 17-21. doi.org/10.1080/00958969909601873

Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology*, **31(2)**, 144-152.

Clason, D. L., & Dormody, T. J. (1994). Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education*, **35(4)**, 32-35.

Conover, W. J. (1973). On methods of handling ties in the Wilcoxon signed-rank test. *Journal of the American Statistical Association*, **68(344)**, 985-988.

Derrick, B. (2017). Package Partiallyoverlapping: Partially overlapping samples t-tests. R package version 1.0.

Derrick, B., Dobson-McKittrick, A., Toher, D., & White P. (2015). Test statistics for comparing two proportions with partially overlapping samples. *Journal of Applied Quantitative Methods,* **10(3)**, 1-14.

Derrick, B., Russ, B., Toher, D., & White P. (2017). Test statistics for the comparison of means for two samples which include both paired observations and independent observations. *Journal of Modern Applied Statistical Methods*, **16(1)**, 137-157. doi.org/10.22237/jmasm/1493597280

Derrick, B., Toher, D., & White, P. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods for Psychology*, **12(1)**, 30-38. doi.org/10.20982/tqmp.12.1.p030

Derrick, B., Toher, D., & White, P. (2017). How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017). *The Quantitative Methods for Psychology,* **13(2)**, 120-126. doi.org/10.20982/tqmp.13.2.p120

Derrick, B., & White, P. (2017) Comparing two samples from an individual Likert question. *International Journal of Mathematics and Statistics*, **18(3)**, 1-13.

Guo, B., & Yuan, Y. (2017). A comparative review of methods for comparing means using partially paired data. *Statistical methods in medical research*, **26(3)**, 1323-1340.

Hodge, D. R., & Gillespie, D. (2003). Phrase completions: An alternative to Likert scales. *Social Work Research*, **27(1)**, 45-55.

Hothorn, K. (2006). A Lego System for Conditional Inference. *The American Statistician*, **60(3)**, 257-263. doi.org/10.1198/000313006X118430

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

Kenney, J. F., & Keeping E.S. (1951) *Mathematics of Statistics*, (2nd ed.), Princeton, NJ: Van Nostrand.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*, **22**.

Maisel, N. C., & Fingerhut, A. W. (2011). California's ban on same-sex marriage: The campaign and its effects on gay, lesbian, and bisexual individuals. *Journal of Social Issues*, **67(2)**, 242-263. doi.org/10.1111/j.1540-4560.2011.01696.x

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, **8(1)**, 3-30. doi.org/10.1145/272991.272995

Norman G (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Science Education Theory and Practice,* **15(5)**, 625–632. doi.org/10.1007/s10459-010-9222-y

Nunnally, J. C. (1978). *Psychometric theory,* (2nd ed.), New York: McGraw-Hill.

Pratt, J. W. (1959). Remarks on zeros and ties in the Wilcoxon signed rank procedures. *Journal of the American Statistical Association*, **54(287)**, 655-667. doi.org/10.1080/01621459.1959.10501526

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Roberson, P. K., Shema, S. J., Mundfrom, D. J., & Holmes, T. M. (1994). Analysis of paired Likert data: how to evaluate change and preference questions. *Family Medicine*, **27(10)**, 671-675.

Sisson, D. V., & Stocker, H. R. (1989). Research corner: analyzing and interpreting Likert-type survey data. *Delta Pi Epsilon Journal*, **31(2)**, 81-85.

Uebersax, J. S. (2006). Likert scales: dispelling the confusion. Statistical Methods for Rater Agreement. Available at: http://john-uebersax.com/stat/likert.htm. Accessed: 08082017.