Powerful and Consistent Analysis of Likert-Type Rating Scales

Maurits Kaptein
Industrial Design
Eindhoven University of Technology, the Netherlands
maurits@mauritskaptein.com

Clifford Nass
Department of Communication
Stanford University, CA, USA.
nass@stanford.edu

Panos Markopoulos
Industrial Design
Eindhoven University of Technology, the Netherlands
p.markopoulos@tue.nl

ABSTRACT

Likert-type scales are used extensively in evaluations of interactive experiences, including usability evaluations, to obtain quantified data regarding participants' attitudes, behaviors, and judgments. Very often this data is analyzed using parametric statistics like the Student t-test or ANOVA. These methods are chosen to ensure higher statistical power of the test (which is necessary in this field of research and practice where sample sizes are often small), or because of the lack of software to handle multi-factorial designs nonparametrically. This paper presents to the HCI audience new developments from the field of medical statistics that enable analyzing multiple factor designs nonparametrically. We demonstrate the necessity of this approach by showing the errors in the parametric treatment of nonparametric data in experiments of the size typically reported in HCI research. We also provide a practical resource for researchers and practitioners who wish to use these new methods.

Author Keywords

Usability evaluation, Nonparametric statistics, Research methods

ACM Classification Keywords

H.5.2 User interfaces: Evaluation/methodology

General Terms

Experimentation, Measurement.

INTRODUCTION

One of the most popular ways to obtain self-report data from participants in evaluations, experiments, and surveys reported in the human-computer interaction literature are *Likert-type*[1] scales [13]. The appropriate analysis of Likert scale measurements – scores computed from a number of relating items – and ratings obtained from Likert-*type* items – *measurements on a single disagree to agree item* – has lead to elaborate discussions [7, 6]. In this article we focus on the analysis of scores on individual Likert-type items and show (1) that parametric methods are not invariant to monotone transformations of this type of data, leading to inconsistent results, and (2) that nonparametric methods exist to fully analyze factorial designs without the generally assumed loss of power leading to a consistent and powerful analysis of the obtained data.

Likert-type rating scales are presented as essential tools in HCI textbooks [15] and are suggested as a very efficient way of collecting self-report data in usability evaluation practice (see e.g., [18]). With the gradual broadening of the focus of this field towards studying and designing for the user experience (see e.g., [11]), or emotional aspects of technology use such as presence, connectedness, etc., the use of Likert-type scales will grow as a way to obtain quantifiable self-report data from study participants, whether these concern experiences, judgments or attitudes.

An examination of the CHI 2009 proceedings shows that 45.6% of all the published articles use Likert-type scales for some measure of the subjective user experience. Within this group 80.6% use parametric tests, such as t-tests or ANOVA, to analyze their data. Only 8.3% of the studies using Likert-type scales report using a nonparametric statistic and the vast majority (94.4%) of studies report on relatively small sample sizes ($N < 50$) for statistical inference.

In this article we present a novel way of analyzing the results obtained from Likert-type scales using newly developed nonparametric statistical methods. We focus on the common 2x2 factorial design: an experiment with two independent variables composed of two levels each. However, this approach can be extended to N-level factorial designs. We show, through simulation, that the nonparametric approach is (1) more powerful for the sample sizes usually found in HCI research than the parametric approach, and (2) is consistent over monotone transformations. As such it makes replication across studies more consistent. We conclude by providing practical resources for researchers and practitioners to conduct the proposed statistical analysis.

[1] We refer to Likert-*type* rating scales since Likert in his original article does not propose using the raw scores on individually rated items as a stand-alone measure, as is frequently done in HCI.

CHI 2010, April 10 – 15, 2010, Atlanta, Georgia, USA

INFERENTIAL STATISTICS AND SIGNIFICANCE

Consider a prototypical HCI experiment comparing user satisfaction with two different operating systems, e.g., Windows Vista vs. Apple Mac OS X. We measure ease of use by employing a 7-point Likert-type scale *"The system was easy to use,"* with response categories Strongly disagree (= 1) to Strongly agree (= 7). We perform this measurement (1) after one week of usage and (2) after one month of usage. We now have a 2x2 mixed design experiment with *type of operating system (A)* as a between-participants factor and *time of measurement (B)* as a within-participants factor. After one month, we have obtained our data and are confronted with the question of how to report the outcomes of our experiment.

For the experiment described above, most practitioners and researchers would use a 2x2 ANOVA, a parametric test, to test the main effects of the two independent variables and a possible interaction effect. This option is commonly chosen because it is viewed as most powerful, it is familiar to most of us, and it is easy to carry out using analysis software.

Data transformations

The validity of the parametric ANOVA depends on a number of assumptions about the collected data. While there is debate among statisticians over the invariability of parametric measures under violations of these assumptions [1, 8, 9, 16, 19] it is clear that parametric methods are not invariant to monotone transformations of the collected data.

A monotone data transformation is any transformation that keeps the order of scores constant while changing the absolute magnitude between the observed data points. As such, a 7-point Likert-type scale can be assigned numerical values 1 to 7, but could also be assigned the numerical values 1, 2, 5, 20, 35, 37, 38 – leaving the order of scores constant.

Researchers use these types of transformations – e.g., log transformations, to adhere to normality demands of the statistical analysis used or to obtain significant results not found in the untransformed data.

Parametric procedures compare mean scores (and distributions around these means) by assigning different numerical values to the obtained measurements thus influences the results. The differing results of parametric methods under monotone transformations raises questions about the validity and reliability of the parametric approach, as it relies on a valid and reliable mapping of scores given by respondents and the value assigned to them for the analysis.

There is little justification to assume that Likert-type scales should be given the standard numerical values 1 to N where N is the number of points of the N-point scale. Although it is clear that the N points represent an order, fixed distances between the points cannot be reasonably assumed. Actually, empirical research has clearly shown that the distances between the points of one single item are *not equal in the perception of respondents* filling out the item [10]. When asked to attribute numerical values to Likert-type scales, respondents report bigger differences between the extremes, the step from 1 to 2 or 6 to 7 on a 7-point scale, as compared to the moderate scores, the step from 3 to 4 and 4 to 5.

Alternative to parametric methods, nonparametric methods depend merely on the order of ratings, and not on the seemingly arbitrary distribution of numerical values. While most researchers and practitioners are aware of the invariance of nonparametric tests under monotone transformation, articles that report a nonparametric approach to statistical analysis are a clear minority in the HCI literature. This is partially caused by the concern of many researchers and practitioners that nonparametric methods are less powerful than parametric methods. However, previous research has shown an increase in power of nonparametric methods over parametric methods for small sample sizes [8] while leading to more consistent and thus easier-to-replicate results. A second reason for the minority usage of nonparametric statistics is the unavailability of software to perform such analysis. This has recently changed with the introduction of SAS macro's and R code to analyze multiple factor designs nonparametrically [2, 17].

THE NONPARAMETRIC APPROACH

Most researchers and practitioners are familiar with the Wilcoxon Signed Ranks test or the Mann–Whitney U test, the nonparametric equivalents of the within and between subjects Student t-tests. Because these are appropriate only for single factor experiments using only two levels for the independent variable, the analysis of multifactor designs usually is done parametrically. Recently, however, several authors in the medical field have extended the nonparametric model to enable analysis of multifactorial designs [2, 3, 4, 14, 17, 17]. In the next section we will introduce the basic underlying concepts of this nonparametric model. This technique is not based on assumptions of population parameters such as the mean and variance. Instead, it is based on effects of variables in reference to the distribution of variables as measured.

Parametric vs. nonparametric hypothesis

The parametric null hypothesis tested in the simplest case – a between-subjects t-test – states that the mean scores of two groups in an experiment are equal. Based on (1) the number of observation in each group, (2) the magnitude of the numerical difference in their means, and (3) the spread of observations around this mean, the probability that the measured result originated by sampling from a population in which there is no difference between the group means is computed. We use this probability to test the null hypothesis that in the population the group means are equal:

$$H_0 : \mu_1 = \mu_2 \text{ or } H_0 : \mu_1 - \mu_2 = 0$$

Contrary to the parametric hypothesis, the nonparametric hypothesis is not based on mean values, nor on assumptions about these parameters in the population. The nonparametric null hypothesis states that the distribution of observed answers over the ordered response categories for one experimental group is equal to the distribution of observed answers in the other experimental group:[2]

$$H_0 : \phi_1(x) = \phi_2(x)$$

As such, it is clear that the nonparametric hypothesis does not depend on arbitrarily-assigned numerical values and thus is invariant under monotone transformations.

Comparisons of experimental groups are based on overall rank scores and rank scores of individual experimental groups. In the Appendix, we give a mathematical overview of the basic approach and references to the original material.

COMPARING METHODS

To test whether indeed the nonparametric method is (1) consistent over monotone transformations, and (2) results in higher power, we simulated the data of the 2x2 experiment which compares the usability ratings of both Windows Vista as well as Mac OS X over two points in time. After simulating responses for a total of 200 experimental cases evenly distributed over the between subject conditions, we performed several monotone transformations on the original ratings which were originally assigned numerical values 1 to 7. Table 1 presents an overview of the numerical values of the transformations.

Name	**Type**	**Transformation**
θ_0	Raw score	1, 2, 3, 4, 5, 6, 7
θ_1	Left skewed	1, 2, 4, 10, 20, 40, 70
θ_2	Right skewed	1, 31, 51, 61, 67, 69, 70
θ_3	Stretched	1, 2, 10, 35, 60, 69, 70
θ_4	Realistic	1, 25, 30, 35, 40, 45, 70

Table 1. Description of the transformations used on the data

To test the effects of these transformations we performed both a 2x2 mixed design ANOVA and a 2x2 nonparametric analysis on the original scores and each of the transformations. To detect sensitivity to sample size, we randomly selected subsets of 40 and 10 experimental subjects, again evenly distributed over the between subject conditions, from the original 200 simulated cases.

Parametric analysis

Table 2 presents an overview of the p-values derived from the 2x2 mixed subjects analysis of variance performed on the $N = 200$, the $N = 40$, and the $N = 10$ simulations. First, it is clear that the ANOVA is not invariant over the monotone transformations: p-values for the different transformations differ greatly. This leads to several opposing conclusions. When looking at the most representative HCI experiment case ($N = 40$), it is clear that based on the original observations, only an effect of time would be reported upon. Given a realistic transformation, in which the extreme answers are numerically more distant than the moderate answers, this difference disappears. When skewing the results (as is effectively done by the frequently-used log transformations) we even find a scenario in which we conclude a statistically significant difference between the usability of Microsoft Vista and of Mac OS X.

$N = 200$	θ_0	θ_1	θ_2	θ_3	θ_4
A (between)	**.000**	**.000**	**.000**	**.000**	**.000**
B (within)	**.000**	**.000**	**.084**	**.000**	**.024**
A*B	.119	.233	.705	**.004**	.160
$N = 40$	θ_0	θ_1	θ_2	θ_3	θ_4
A (between)	.055	**.008**	.461	**.030**	.084
B (within)	**.018**	**.002**	.794	**.001**	.795
A*B	.662	.199	.271	.092	.849
$N = 10$	θ_0	θ_1	θ_2	θ_3	θ_4
A (between)	.320	.596	.243	.198	.673
B (within)	.100	**.039**	.145	.052	.057
A*B	.645	.869	.728	.905	.419

Table 2. Outcomes in p-values of a 2x2 mixed subjects analysis of variance on the simulated data. (A = type of operating system, between-participants and B = time, within-participants)

Nonparametric analysis

Table 3 presents the outcomes in p-values using the Anova Type Statistic (ATS - see Appendix) for the nonparametric approach as computed using the SAS macros provided by Brunner [2]. The results are invariant to the monotone transformations and thus no opposing conclusions can be drawn based on the transformations. Equally important, however, is the increased power – depicted in the on-average decreased p-values. Thus, experimental differences are more easily and more consistently identified using the nonparametric approach. Based on the nonparametric approach we would confidently report a main effect of type of operating system, and a main effect of time.

	$N = 200$	$N = 40$	$N = 10$
A (between)	**.000**	**.028**	.394
B (within)	**.000**	**.000**	**.001**
A*B	**.025**	.304	.426

Table 3. Outcomes in p-values of a 2x2 mixed subjects analysis of nonparametric analysis on the simulated data. Obtained p-values are equal for θ_0 to θ_4; The nonparametric method is invariant to the monotone transformation.

CONCLUSIONS

In this note, we introduced the concepts of the general N-factor nonparametric approach to the analysis of Likert-type scales by comparing the results of a 2x2 mixed design analyzed both parametrically as well as nonparametrically. The nonparametric approach presents a recently developed analysis method that is not yet used within the HCI field. Since Likert-type scales are frequently used in HCI, and relatively low sample sizes ($N < 50$) are common, we advocate the usage of this nonparametric approach in favor of the general parametric ANOVA. While usually ignored because of their lack of power we have shown that in small sample size situations, common in HCI, nonparametric approaches can lead to a power increase.[3] We also argue that

[2] We use $\phi_1(x)$ instead of the $F_1(x)$ used by Brunner [2] to avoid confusion with the F-statistic.

[3] A more elaborate proof of the power increase of nonparametric methods over parametric methods in the case of small samples sizes can be found in [5].

the second major reason of rejection of such approaches, the unavailability of software to conduct the analysis, has been addressed. On the website accompanying this CHI note (www.nth-iteration.com/study/statistics) we have grouped both the available SAS macros and the R-code for several experimental designs, and provide an online tool to analyze 2x2 mixed subject designs.

REFERENCES

1. C. A. Boneau. The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, 57:49–64, 1960.
2. E. Brunner, S. Domhof, and F. Langer. *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*, volume Total Quality Management. John Wiley & Sons, 2002.
3. E. Brunner and U. Munzel. *Nichtparametrische Datenanalyse*. Springer, 2002.
4. E. Brunner, U. Munzel, and M. L. Puri. Rank-score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis*, 70:286–317, 1999.
5. E. Brunner and M. L. Puri. Nonparametric methods in factorial designs. *Statistical Papers*, 42:1–51, 2001.
6. J. Carifio and R. Perla. Resolving the 50-year debate around using and misusing Likert scales. *Med Educ*, 42:1150–52.
7. J. Carifio and R. Perla. Ten Common Misunderstandings, Misconceptions, Persistent Myths and Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences*, 3:106–116.
8. R. A. Cribbie and H. J. Keselman. The effects of nonnormality on parametric, nonparametric, and model comparison approaches to pairwise comparisons. *Educational and Psychological Measurement*, 63:615–635, 2003.
9. G. V. Glass, P. D. Peckham, and J. R. Sanders. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3):237–288, 1972.
10. M. C. Hart. *Improving the discrimination of SERVQUAL by using magnitude scaling*, volume Total Quality Management. Chapman & Hall, 1996.
11. M. Hassenzahl. The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction*, 19:319–349, 2004.
12. S. Hora and W. Conover. The f-statistic in the two-way layout with rank-score transformed data. *Journal of the American Statistical Association*, 79:668–673, 1984.
13. R. Likert. A technique for the measurement of attitudes. *Journal of Social Psychology*, 5:228–238, 1932.
14. U. Munzel and B. Bandelow. The use of parametric versus nonparametric tests in the statistical evaluation of ratings scales. *Pharmacopsychiatry*, 6:222–224, 1998.
15. J. Preece, Y. Rogers, and Sharp. *Interaction Design*. John Wiley & Sons, 2002.
16. S. S. Sawilowsky and R. C. Blair. A more realistic look at the robustness and type ii error probabilities of the t test to departures from population normality. *Psychological Bulletin*, 111:352–360, 1992.
17. D. A. Shah and L. V. Madden. Nonparametric analysis of ordinal data in designed factorial experiments. *Pythopathology*, 94:33–60, 2004.
18. T. Tullis and W. Albert. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufman, 2008.
19. R. R. Wilcos. Comparing the means of two independent groups. *Biometrics Journal*, 32:771–780, 1990.

APPENDIX

One-way nonparametric approach

The rating of one individual on a Likert-type scale can be denoted as X_{ik}, which is a random variable with normalized distribution $\phi_i(x)$. From here on we can proceed by computing a weighted average of all the $\phi_i(x)$'s in an experiment. Given 2 values of i – two experimental groups – and 7 values of x – seven points on the Likert-type scale -, the weighted average is given by:

$$H(x) = \frac{1}{N} \sum_{i=1}^{2} n_i \phi_i(x)$$

in which N is the total number of respondents. From this weighted average it is clear that more weight is given to groups with more observations and as such are more important in determining the average distribution.

Now, given $\phi_i(x)$ and $H(x)$ we can define the effect of the ith treatment as:

$$p_i = \int H d\phi_i$$

in which $d\phi_i$ is the first derivative of ϕ_i. The p_i value describes the stochastic tendency of ϕ_i with respect to H. If $p_i > \frac{1}{2}$ observations in the ith treatment tend to be larger than an independent random variable which has H as its distribution. Thus, ratings on this item are more positive for the ith experimental group than for the overall sample.

Direct calculations of p_i are quite tedious [17, 2]. However, it can be shown that the estimated relative treatment effect $\hat{p_i}$ can be determined directly from observation midranks. Midranks are ranks corrected for possible ties. Based on the estimated treatment effects, one can compute the ANOVA type statistic (ATS) [12, 2] which, based on asymptotic theory, has an approximate F distribution under the null hypothesis and is especially suitable for small sample sizes ($10 < N < 200$).

N-Factorial designs

The approach presented above can be extended to N-Factorial multilevel experiments. When considering a levels of factor **A** and b levels of factor **B**, the weighted average distribution can be computed using:

$$H = \frac{1}{N} \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} \phi_{ij}$$

The relative effect of the ith level of **A** and the jth level of **B** is then given by:

$$p_{ij} = \int H d\phi_{ij}$$

Again, midranks can be used for a direct estimation of treatment effect size based on the observations and the ATS statistic can be used to test both the two main effect null hypotheses and as well as the possible interaction effect. The null hypothesis for the main effect is a straightforward extension of the one-way two level null hypothesis: $H_0(\mathbf{A}) : \overline{\phi}_{1\cdot} = \overline{\phi}_{2\cdot} = \overline{\phi}_{3\cdot} = \ldots = \overline{\phi}_{a\cdot}$.The null hypothesis for the interaction effect is given by: $H_0(\mathbf{A*B}) : \overline{\phi}_{ij} + \overline{\phi}_{\cdot\cdot} = \overline{\phi}_{i\cdot} + \overline{\phi}_{\cdot j}$.