Appraisal of Statistical Practices in HRI vis-á-vis the T-Test for Likert Items/Scales

Matthew Gombolay* **and Ankit Shah***
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

Abstract

Likert items and scales are often used in human subject studies to measure subjective responses of subjects to the treatment levels. In the field of human-robot interaction (HRI), with few widely accepted quantitative metrics, researchers often rely on Likert items and scales to evaluate their systems. However, there is a debate on what is the best statistical method to evaluate the differences between experimental treatments based on Likert item or scale responses. Likert responses are ordinal and not interval, meaning, the differences between consecutive responses to a Likert item are not equally spaced quantitatively. Hence, parametric tests like t-test, which require interval and normally distributed data, are often claimed to be statistically unsound in evaluating Likert response data. The statistical purist would use non-parametric tests, such as the Mann-Whitney U test, to evaluate the differences in ordinal datasets; however, non-parametric tests sacrifice the sensitivity in detecting differences a more conservative specificity – or false positive rate. Finally, it is common practice in the field of HRI to sum up similar individual Likert items to form a Likert scale and use the t-test or ANOVA on the scale seeking the refuge of the central limit theorem. In this paper, we empirically evaluate the validity of the t-test vs. the Mann-Whitney U test for Likert items and scales. We conduct our investigation via Monte Carlo simulation to quantify sensitivity and specificity of the tests.

Introduction

Human-robot interaction (HRI) is a broad, interdisciplinary field bringing together researchers in psychology, sociology, anthropology as well as in computer science, electrical engineering and mechanical engineering. Together, these researchers study the interaction of humans and robots. Research can typically be divided into two general types. The first type of research focuses on the social science aspects of interaction between humans and robots. For example, a recent paper investigated public perceptions of sex robots (Scheutz and Arnold 2016). The second type of research focuses on the engineering aspects of the interaction. For example, a paper that received a best-paper award for "enabling algorithms" employed a Mixed-Observable Markov Decision Process and Inverse Reinforcement Learning to give a robot the ability to learn from demonstration how to ergonomically position a surface for a human to paint (Nikolaidis et al. 2015).

We surveyed the proceedings of HRI 2015 (HRI'15), which consisted of 43 accepted papers. The acceptance rate for this conference is roughly $\sim$ 30%. Of these 43 accepted papers, 40 of 43 (93%) involved human-subject experimentation. The other three were papers focused on statistical modeling rather than human subject experimentation. Of these 40 papers, 23 papers (58%) used a Likert-type response to solicit perceptions of human subjects as a function of one or more experimental variables.

A Likert item is an ordinal scale to solicit an experimental participants level of agreement with a statement regarding an experimental condition they experience. As such, best practices dictate that one should employ a non-parametric test, such as a Mann-Whitney U test to asses differences. Use of a t-test, z-test, or ANOVA assumes interval data and that the residuals are normally distributed. It is plausible that one could argue via the Central Limit Theorem (CLT) that summing across multiple Likert items in forming a Likert scale provides an aggregate measure that can roughly assumed to be normally distributed. However, in our review of the proceedings for HRI'16, of the 23 papers that use a Likert-response format to solicit perceptions from subjects, 16 (70%) applied a statistical test to individual response item. Thus, any hope for leveraging the CLT is lost. Furthermore, 21 (90%) applied a t-test or analysis of variance (ANOVA) – both of which assume the data are interval and normally distributed. Not a single one of these papers first used a test for normality, such as a Chi-Squared Goodness-of-Fit test.

Where does this leave us? Should we ignore all of the results in HRI'15? These are the questions that we seek to answer in the remainder of this paper. Through an extensive set of Monte Carlo Simulations, we find that, in fact, the papers that employ a t-test, even for single Likert items, are likely valid. Specifically, we find that the false positive rate is empirically less than or equal to the significance level α. However, we also caution against the ubiquitous practice of testing individual Likert items: The chance of falsely finding a positive results increases exponentially.[1]

*These authors contributed equally to the work.

[1] The probability of a false positive result is $Pr\{$False

First, we start with an overview of Likert items and scales, the central limit theorem, best practices for design of these response formats and statistical testing. Next, we present our experimental methods and design of our Monte Carlo simulations to test the validity of common statistical practices in HRI. We then present the results of our Monte Carlo simulations. We discuss the implications of our findings and present a set of guidelines for authors, reviewers, and the HRI community at large. Next, we present the limitations of our work and proposed, future work.

Preliminaries

In this section, we first introduce the notion of a Likert item, which can then be used to construct a Likert scale. We present best practices in constructing these items and scales as well as common pitfalls in their use. Finally, we present two approaches – non-parametric and parametric – to infer differences in factor level medians of ratings on a Likert item or scale in the context of a single-factor, two factor level experimental design.

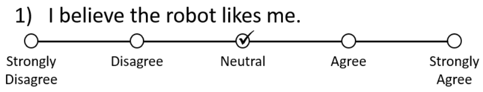
Likert Item

Consider a scenario where a roboticist is designing behaviors for a robot to make it easier for the robot to work with a human on a joint task. The roboticist wants to know which of the behaviors works best to facilitate the interaction. As a first thought, the roboticist conducts an experiment with a sample population, perhaps a small group of potential users. The experimenter divides the sample population into two groups: groups A and B. Group A experiences one behavior of the robot and group B experiences a different robot behavior. To solicit the users' views, the roboticist shows them with a prompt (Figure 1), which states, "I believe the robot likes me." The participants are asked to rate the degree to which they agree or disagree with the statement. The roboticist establishes a null and alternate hypothesis as follows:

- H_0 - There is no difference in the average responses of the participants between groups A and B.
- H_1 - There is a difference in the average responses of participants between groups A and B.

Finally, the experimenter would apply the appropriate statistical test and draw conclusion based on that test, looking at the data in aggregate.

This scenario is analogous to numerous experimental scenarios that industry practitioners and academic researchers construct to answer important questions in their practice. In this particular scenario, the experimenter chose an experimental design with one factor (behavior mode), two factor levels (behavior A and B), in which subjects each experienced only one factor level. The method of response employed in our example scenario is known as a *Likert item*. A Likert item is a statement that a human subject (i.e., an experimental unit) is asked to evaluate along a subjective or objective dimension. Most commonly, a Likert item gives a statement, such as "I believe the robot likes me," and the user is asked to evaluate his or her level of agreement or disagreement along a response scale shown below the statement. However, Likert items can take other forms. For example, the prompt can simply be to rate how much pain you are experiencing, and the subject should select the "feeling" that best describes how they feel, as shown in Figure 2. Because the response format for a Likert item is often presented as a scale, people often refer to a single Likert item as a Likert scale, which is a misnomer.


1) I believe the robot likes me.

○ Strongly Disagree ○ Disagree ☑ Neutral ○ Agree ○ Strongly Agree

Figure 1: This figure depicts an example of a Likert item for rating a user interface (Hoffman 2013).

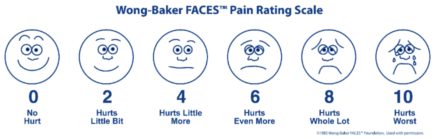

Wong-Baker FACES™ Pain Rating Scale
0
No Hurt
2
Hurts Little Bit
4
Hurts Little More
6
Hurts Even More
8
Hurts Whole Lot
10
Hurts Worst
©1983 Wong-Baker FACES® Foundation. Used with permission.

Figure 2: This figure depicts an example of a Likert item for rating pain or discomfort.

Statistical Testing for Likert Items

Data from Likert items are inherently ordinal: While one can know that a rating of "strongly agree" is greater than "strongly disagree," one cannot say that a change from "agree" to "strongly agree"is greater than a change from "disagree" to "neutral." One cannot assess the *distance* between points along the response scale; however, one can create a ranking of responses. As such, the data are ordinal. With ordinal data, one cannot assume the data are normally distributed, therefore, mathematically, we cannot assume that response x_i comes from a normal distribution $x_i \sim \mathcal{N}(\mu, \sigma^2)$).

$$t_{2n-2} = \frac{\bar{x}_A - \bar{x}_B}{s_{x_A x_B}\sqrt{\frac{1}{n}}} \tag{1}$$

$$s_{x_A x_B} = \sqrt{s_{x_A}^2 + s_{x_B}^2} \tag{2}$$

$$s_{x_i} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{i,j} - \bar{x}_i)^2} \tag{3}$$

If one were given this assumption of normality, one could employ a Student's t-test, which leverages this normality assumption, to determine the likelihood p that differences in factor level means are due to random noise. The t-test is defined in equation 1, where $\bar{X}_A$ and $\bar{X}_B$ are the mean responses of the Likert item for subject groups A and B. The pool sample standard deviation $s_{x_A x_B}$ is shown in Equation 2, and the group standard deviations s_a and s_b are shown in Equation 3. Finally, n is the number of subjects in each

Positive$\} = 1 - (1-\alpha)^n$, where n is the number of individual items tested, and α confidence level below which a result is assumed positive.

group. This formulation of the t-test assumes both groups have an equal variances and an equal number of subjects n. The subscript for t shows the degrees of freedom of the test. The degrees of freedom for this test are $2n - 2$ because we have have have $2n$ subjects and lose two degree from the estimation of the factor level means $\bar{x}_A$ and $\bar{x}_B$.

After calculating the t-statistic, t_{2n-2}, through Equation 1, one can look up the corresponding p-value given the degrees of freedom. This p-value is the probability that any difference in the factor level means is due to random noise (i.e., a Type I Error). if p is less than some confidence level α, then we say that we reject the null hypothesis that there are no differences in the mean response between the two factor levels (i.e., web page interfaces). Commonly, α is set to 0.05, meaning that the experimenter is willing to accept a 5% chance that a perceived difference in the results is actually erroneous. Such an error is known as a Type I error and is typically set at $\alpha = 0.05$.

Alas, we are not given the assumption of normality. Instead, one must use a non-parametric test, such as a Mann-Whitney U test, to infer whether differences in the two experimental groups exist based on their subject ratings obtained through the single, Likert item. This test relies on evaluating the ranks of responses between the groups rather than the value of the responses. The test is formulated, as shown in Equations 4-8. Here, R_i is the sum of the overall rank of the j^{th} subject's response to the Likert item from group i (Equation 4). U_i is the a measure of this rank for group i, and U is the lesser of U_1 and U_2 from groups 1 and 2 (or A and B in our web page interface example).

$$R_i = \sum_{j=1}^{n_i} r_i \tag{4}$$

$$U_i = R_i - \frac{n_i(n_i+1)}{2} \tag{5}$$

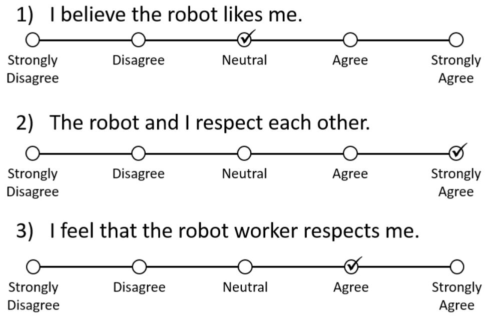
$$U = \min\left(U_1, U_2\right) \tag{6}$$

$$z = \frac{U - \mu_U}{\sigma_U} \tag{7}$$

$$\mu_U = \frac{n_1 n_2}{2} \tag{8}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 \left(n_1 + n_2 + 1\right)}{12}} \tag{9}$$

Finally, we can formulate our test statistic. As the sample size approaches infinity, U approaches a normal distribution. As such, we can test whether there is a difference in the factor level medians according to Equation 7, where μ_U is the mean under the null hypothesis that there is no difference in the factor level medians between the responses (Equation 8), and σ_U is the sample standard deviation of U under this null hypothesis (Equation 9). This formulation of the Mann-Whitney U test assumes that there are no ties amongst the ranks. If ties exist, there is a correction term for σ_U, which can be used instead [2]



1) I believe the robot likes me.

○ Strongly Disagree ○ Disagree [✓ Neutral] ○ Agree ○ Strongly Agree

2) The robot and I respect each other.

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree [✓ Strongly Agree]

3) I feel that the robot worker respects me.

○ Strongly Disagree ○ Disagree ○ Neutral [✓ Agree] ○ Strongly Agree

Figure 3: This figure depicts an example of a Likert scale (Hoffman 2013).

Unfortunately for practitioners, non-parametric tests, such as the Mann-Whitney U test, are typically less sensitive, meaning that the likelihood of rejecting the null hypothesis, given that the null hypothesis is incorrect, is less than for a parametric test. It is tempting, then, to want to apply a parametric test, such as a t-test, for the sake of finding a statistically significant result. However, this is theoretically unsound. As such, what is a sound practitioner supposed to do? The answer, according to psychologist and inventor Rensis Likert, is a Likert Scale.

Likert Scale

A Likert scale is a summation across a set of Likert items. Figure 3 shows an example for our robot behavior design experiment. When designing a Likert scale, it is best to follow two principles. First, each Likert item response scale should be symmetric. The Likert item shown in Figure 3 is symmetric. An example of an asymmetric scale would be if the most positive rating was "strongly agree" and the most negative rating was "neutral." Second, a Likert scale should be balanced: The scale should have an equal number of positive and negative prompts. In the example shown in Figure 3, the principle of symmetry is followed, but the principle of balance is not. There are three positive statements and zero negative statements. It would be better to make the last prompt read "3. I feel the robot worker does not respect me."

Likert scales also have pitfalls. First, experiment participants responding to a Likert scale are likely to respond with a less-extreme position, for example, by responding with "agree" when they may, in fact, "strongly agree." It is well known that subjects are subject to a *central tendency bias*, meaning that they do not want to appear extreme or different from the average person. Second, subjects are likely to agree with the prompts given. This bias is known as the *acquiescence bias*. Other pitfalls are common across many response types, such as the *experimenter-expectancy effect*, in which subjects try to respond in the way they think the experimenter desires.

Statistical Testing for Likert Scales

A Likert Scale is a powerful technique for measuring a subjective perception regarding the effect of a treatment or factor because it can take advantage of the CLT. Let us assume

[2] $\sigma_{corr} = \frac{n_1 n_2}{12}\left((N+1) - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{N(N-1)}\right)$, where N is the total number of subjects $n_1 + n_2$, t_i is the number of subjects sharing i^{th} rank, and k is the number of tied ranks.

that we have a function that maps each response along the scale of the Likert item (“strongly agree” to “strongly disagree”) to a numeric value. An example coding could be as follows:

- “Strongly Agree” → 1
- “Agree” → 2
- “Neutral” →3
- “Disagree” → 4
- “Strongly Disagree” → 5

More technically, we have a function mapping $f : l_i \rightarrow x_i$, where function f maps Likert response i, $l_i \in$ {“Strongly Agree”,“Agree”,“Neutral”,“Disagree”,“Strongly Disagree”}[3] to $x_i \in \mathbb{R}$. Let us assume this mapping is monotonically non-decreasing. The idea of the CLT was first presented by which was orginally developed in (Erdös and Kac 1946) and (Donsker 1951) and developed by the likes of Billingsley, Prohorov, Skoroh, and more (Brown and others 1971). The CLT states that, as the number of samples x_i increase (e.g., number of items in the Likert scale increases), the distribution of the responses approaches a normal distribution with mean of zero and variance σ^2. The CLT is shown in Equation 10. Furthermore, it has also been shown that $\sqrt{n}(S_n - \mu)$ approaches a normal distribution with mean 0 and variance σ^2, as shown in Equation 11.

$$\lim_{n\to\infty} S := \lim_{n\to\infty} \frac{x_1 + \ldots + x_n}{n} \rightarrow E[x_i] = \mu \tag{10}$$

$$\lim_{n\to\infty} \sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \mu\right) \rightarrow \mathcal{N}(0, \sigma^2) \tag{11}$$

These equations hold true for random variables that are independent and identically distributed, with mean $E[x_i] = \mu$ and finite variance $var[x_i] = E[x_i^2] - E[x_i]^2 < \infty$.

What is incredibly powerful is that we can leverage this theorem to use a parametric test (i.e., a z-test) to test for differences in subjects’ responses to Likert scales as a function of factor level with one catch: the number of items in the Likert scale needs to approach infinity. We could argue that we can correct for this assumption by using a t-test, which accounts for small sample sizes. However, a t-test still relies on data that are at least interval, meaning that one can measure, quantitatively, the distances between two responses on the response scale. However, Likert items are not inherently interval. We must then ask two questions:

1. Does a Likert item sufficiently approximate interval data such that a t-test can be used?
2. Is a t-test robust to small Likert scales?

In the remainder of this paper, we conduct a Monte Carlo experiment to answer these questions. We begin with a description of the methods for our computational investigation.

Methods

In this section, we describe our computational experiment to test the robustness of the t-test in scenarios where non-interval data is treated as normally distributed interval data. We perform a hypothetical between subjects experiment with a single factor with two treatment levels. The responses to Likert items are generated by virtual subjects, and the two data sets generated are compared using the t-test and the Mann-Whitney U test. The tests are replicated numerous times to measure the relative performance of the two tests under various conditions. Here, we first describe the dependent variables are their levels considered for the experiment, the response variables we measure to compare the tests and the methodology used to conduct the experiment

Controlled Factors

We hypothesize that the difference between the tests would depend on the following variables:

1) Response mapping - A Likert item is an ordinal type, where the differences between equally spaced values on the Likert response do not correspond to an equally spaced differences on an actual response scale. We hypothesize that an actual response scale exists, and we map response values generated on that scale back to a Likert response. We consider four mappings for this project. These mappings are defined by the function $f(x) : x \in \{1, 2...7\} \rightarrow \mathbb{R}$, where $x \in \{1, 2...7\}$ is a Likert response and $f(x)$ is the value of a hypothetical “actual” response. Each mapping must be monotonically non-decreasing i.e. $f(x_1) \geq f(x_2)$ if $x_1 > x_2$ and antisymmetric about the neutral point i.e. $f(x_n + x) = -f(x_n - 4)$, where $x_n = 4$ is the neutral Likert response. The first such mapping is the linear map given by Equation 12, where an equal difference in the Likert response corresponds to an equal difference on the actual response scale. We also consider a sigmoid mapping defined in Equation 13, where the differences in the actual response are decreasing towards the extreme end of the scale. Next, we consider the cubic mapping defined in Equation 14, where the differences between the actual response increase towards the extremes of the scale, and are flat close to the center of the scale. Finally, we consider a fifth order polynomial defined in Equation 15, where the differences at the extremes are further exaggerated, and the mapping is flatter close to the neutral point.

$$f(x) = \frac{1}{6}(x - 4) \tag{12}$$

$$f(x) = \frac{1}{1 + e^{-(x-4)}} \tag{13}$$

$$f(x) = (x - 4)^3 \tag{14}$$

$$f(x) = (x - 4)^5 \tag{15}$$

Figure 4 shows the mappings from Likert response to the hypothetical actual responses in each of the described cases.

2) Response mean positions - The responses are generated synthetically on a hypothesized ”actual” response curve and transformed to a Likert response. The positions

[3] The set of possible values depends on the specific likert item. The set commonly has 5 items, but can also include 7, 9, or more.

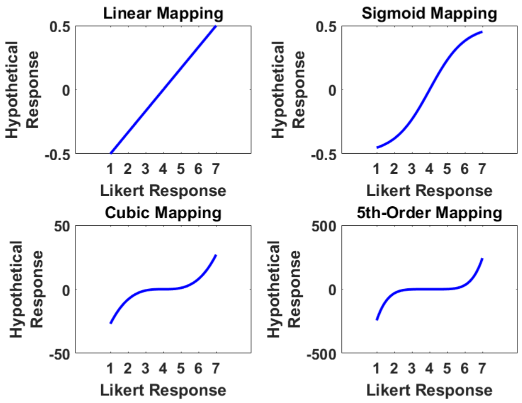


Figure 4: Mappings from actual response to Likert response considered

of the means of the distribution from which the data is sampled are important factor in the error rates of the tests. The means were placed uniformly at six positions between the extremes of the Likert response scale. The positions were $\mu_1, \mu_2 = \{1, 2.2, 3.4, 4.6, 5.8, 7\}$. While sampling the population means were placed at $f(\mu_1)$ and $f(\mu_2)$ respectively.
3) Number of respondents - As the number of Likert responses summed up to compute the mean increases, the distribution of the mean would tend toward a normal distribution as per the central limit theorem. Due to this, we consider the number of respondents as an important independent variable in the experiment. We consider three cases, where the number of respondents are $n_{respondents} = \{5, 15, 30$ These values are typical of the number of participants in HRI studies. Certain studies where access to experimental subjects is limited, the studies are small scale with 5 to 10 participants. Pilot studies, which are often of an exploratory nature recruit 10 to 20 participants. Larger scale studies recruit greater than 30 participants.
4) Number of Likert items per scale - As described in subsection the responses to multiple questions intended to measure similar subjective measures are coded and summed to generate a Likert scale. As more number of items are added to form the scale, the distribution of their sums tends to normal according to the central limit theorem. We consider three cases where we construct a scale from different number of responses, where $n_{items} = \{1, 5, 10\}$. The case of $n_{items} = 1$ corresponds to the case where a Likert item is individually being compared. The responses to the items comprising the scale for a given treatment are drawn from identical distributions.

Factors held constant

The following factors may affect the outcomes, but for the scope of this project, they were held constant. They are:
1) Sampling Variance - The data is generated by sampling from a normal distribution on the hypothetical "actual" response scale. The variance of the data normalized by the range of the scale is held constant i.e. $\sigma^2 = C/(f_{max} - f_{min})$, where C is held constant, in this project at $C = 0.33$. This helps us obtain an even spread in the Likert responses irrespective of the range of the actual scale. In future work, we will vary σ^2.
2) Number of test replicates - Monte Carlo methods rely on replicating an experiment multiple times to sample a wide range of outcomes, the response variables tend to a limiting value as the replicates tend to infinity. However, in practice it is impossible to generate infinite replicates. Here we conduct $n_{test} = 1000$ replicate tests for each treatment.

Test Method

The following testing method was used

1. For each treatment level:
   (a) For each test:
      i. Sample responses for each Likert item from $\mathcal{N}(f(\mu_i), \sigma^2)$ for both treatments.
      ii. Use the inverse map to generate Likert responses using the mapping function.
      iii. Sum up the responses to generate the Likert Scale.
      iv. Perform a t-test on the Likert scale data at $\alpha = 0.05$ and record the outcome and p-value.
      v. Perform a Mann-Whitney U test on the Likert scale data at $\alpha = 0.05$ and record the outcome and the p-value.
   (b) Compute the average p-value for the each test.
   (c) Generate the contingency table for the data.

Considering that there are six mean locations for two factor respectively, three levels for number of respondents and number of items in a scale, and four mappings, we have 1,296 such experimental conditions.

We also conducted an additional Monte Carlo simulation to determine the statistical power of the respective tests. For this, the means of the distribution from which the responses were sampled were placed symmetrically around the neutral point of the Likert response scale. This simulation had a denser grid of the means with $\mu_1 - 4 = -(\mu_2 - 4)$; $\mu_2 \in \{4, 4.3, 4.6, 4.9, 5.2, 5.5, 5.8, 6.1, 6.4, 6.7, 7\}$ and the means of the distribution were placed at $f(\mu_1)$ and $f(\mu_2)$ respectively. This study was replicated for all values of the other factor levels. This simulation had a total of 396 experimental units.

Data collection and analysis

For each experimental unit, the average p-values for the Mann-Whitney U test and the Student's t-test were recorded. In addition to the average p-values, for each treatment, including a value of μ_1, μ_2, $n_{respondents}$ and n_{items}, a contingency table was generated as shown in Table 1. The contingency tables were analyzed using a χ^2 contingency test to check for statistically significant differences in the error rates for the two tests. For each value of $n_{respondents}$ and n_{items}, the 6×6 matrix of the p-values of the χ^2 test on contingency tables were plotted as a heat map as depicted in Figure 7. For each value of $n_{respondents}$ and n_{items}, the confusion matrix depicted in Table 2 was generated for each of the tests.

In addition, we also measure the false positive rates for both the tests for all values of $n_{respondents}$, n_{items} in the cases where $\mu_1 = \mu_2$. The false positive rates plotted against the position of the means are depicted in figure 5. Finally, the sensitivity of the tests is compared by measuring the average p-value generated by the two tests in the second experiment simulation described in subsection . The average p-values plotted against the difference in means symmetrically placed around the neutral response are depicted in figure 6.

	t-test	Mann-Whitney U Test
Correct		
Incorrect		

Table 1: Contingency table for test error rates

		Predicted	
		Same	Different
Actual	Same		
	Different		

Table 2: Confusion matrix for a given test

Results

In this section, we present a subset of the results of our extensive Monte Carlo simulations. A full set is provided in Appendix I. The full paper with appendices are provided online at http://tiny.cc/jdlldy.

False Positive Rate

Figure 5 depicts the false positive rates for the t-test and Mann-Whitney U test for the extreme case for an experiment with one factor, each factor level contains only five subjects, and each subject responds to only one, 7-point Likert item. We set the significance threshold for rejecting the null hypothesis at $\alpha = 0.05$, which means that the probability of a false positive (i.e., incorrectly rejecting a null hypothesis) is 5%. If the t-tests assumptions are violated, as they are for testing a single Likert item, one would fear that the test to have a high false positive rate. Fascinatingly, the false positive for the t-test is actually at or below ~ 0.05 for all mappings tested. As expected, the false positive rate for the Mann-Whitney U test is less than that of the t-test. The evidence supports the notion that the Mann-Whitney U test is, in fact, overly conservative for the mappings we tested.

Sensitivity (P-Values)

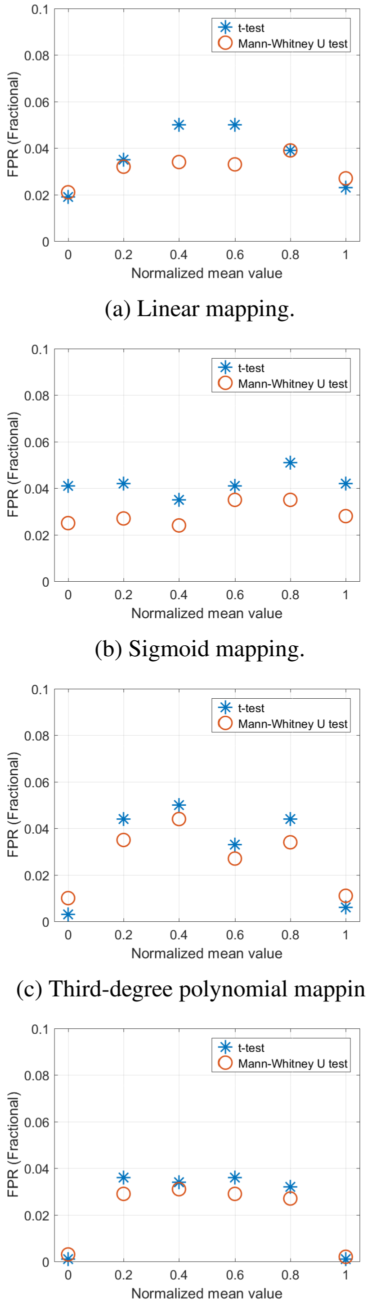
Figure 6 depicts the average p-values for the t-test and Mann-Whitney U test for the extreme case for an experiment with one factor, each factor level contains only five subjects, and each subject responds to only one, 7-point Likert item. We set the significance threshold for rejecting the null hypothesis at $\alpha = 0.05$. Because the t-test makes stronger assumptions about the distribution of the data than the Mann-Whitney U test, we would expect the average p-value for experiments with differences in the factor level means to be lower (i.e., the test is more likely to reject the null hypothesis). We do, in fact, see evidence that the t-test is more sensitive than the Mann-Whitney U test. These figures confirm our intuition. Coupled with the results from the investigation of the false positive rate, we are building a compelling story that the t-test might quite robust to measuring differences in Likert items.

Differences Between the T-test and Mann-Whitney U Test

While we have already seen evidence that difference may exist between the t-test and Mann-Whitney U test, we sought to statistically evaluate this hypothetical difference. Figure 7 depicts the results of this investigation. In each of the plots in Figure 7, the x-axis represents the factor level mean for factor level 1, and the y-axis represents the factor level mean for factor level 2. We measure the aggregate number of correct and incorrect rejections of the null hypothesis for the t-test and Mann-Whitney U test for the extreme case of an experiment with one factor, each factor level contains only five subjects, and each subject responds to only one, 7-point Likert item. Across all four mappings, we find that there is a high-probability of differences in the proportions of correctly and incorrectly rejecting the null hypothesis for these two tests. This data confirm our intuition that the t-test is producing different results than the Mann-Whitney U test. However, the story does not end there. We repeat this analysis but for a hypothetical experiment with 30 subjects per factor level and a Likert scale comprised of 10 Likert items (Figure 8). What we find is that the probability of differences, as measured by a Chi-squared test for independence, largely disappear. This result is surprising, yet makes sense considering the CLT. The results of hypothesis testing for parametric and non-parametric tests seem to converge with increasing replicates and size of the Likert scale despite the ordinality and non-normality of Likert item responses.

Discussion

In our results section, we provide evidence that the t-test is quite robust for testing for the existence of differences between responses measured by a single Likert item. Our findings are, in fact, supported by prior work (Carifio and Perla 2007). First, (Glass, Peckham, and Sanders 1972) showed that the F-test is quite robust to deviations from normality. We know that the F-test and t-test are equivalent for a single-factor experiment with two factor levels. Further, we know that the Likert scale can reasonably be approximated as an interval scale even though it is inherently ordinal (Carifio 1976; 1978). Based on our analysis, we believe that using a t-test for testing Likert items and scales is a reasonably safe practice with a low false positive rate. Further, as the sample size of the experiment increases, the t-test and Mann-Whitney U test are empirically near equivalent. However, we emphatically warn that testing multiple, individual Likert items rather than a single Likert scale greatly increases the chance of a false positive rate. Such practices should be discouraged unless the statistician properly controls for the



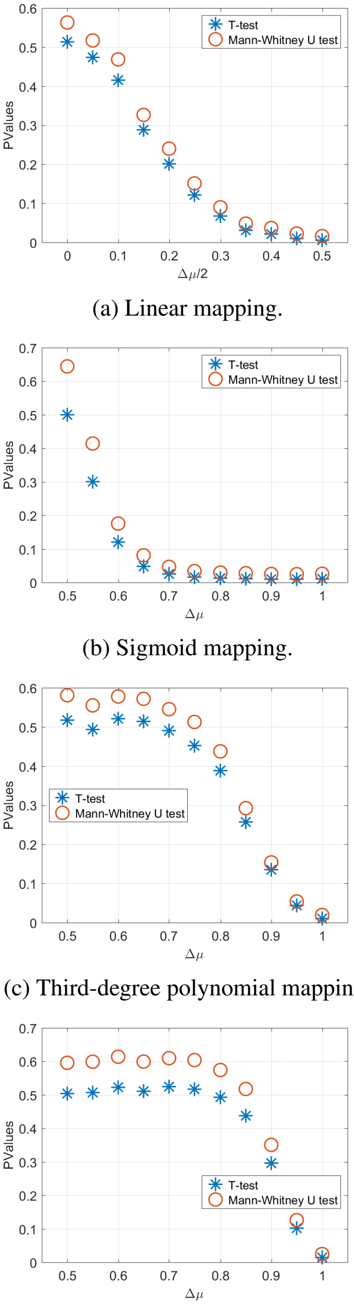
(a) Linear mapping.

(b) Sigmoid mapping.

(c) Third-degree polynomial mapping.

(d) Fifth-degree polynomial mapping.

Figure 5: These figures depict the false positive rates for the Mann-Whitney U test and the t-test when the significance level is $\alpha = 0.05$ for experiments with 5 subjects in each of two factor levels and a Likert scale of only 1 Likert item. Linear, sigmoid, 3^{rd}-degree and 5^{th}-degree polynomials are shown. For these mappings, the x-axis depicts the normalized value of the factor level means given that $\mu_1 = \mu_2 = x$. For $x = 0$, the mean is "strongly disagree", and, for $x = 1$, the mean is "strongly agree."



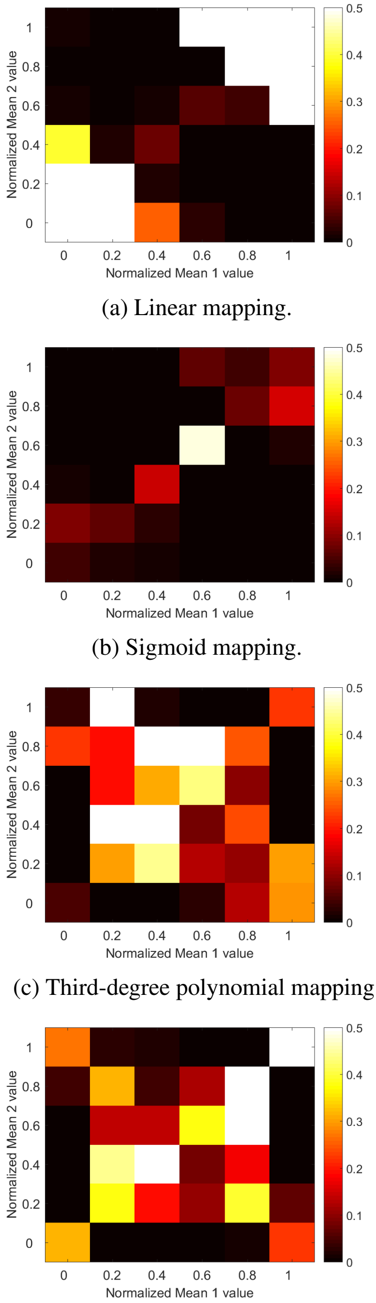
(a) Linear mapping.

(b) Sigmoid mapping.

(c) Third-degree polynomial mapping.

(d) Fifth-degree mapping.

Figure 6: These figures depict the average p-values for the Mann-Whitney U test and the t-test when the significance level is $\alpha = 0.05$ for experiments with 5 subjects in each of two factor levels and a Likert scale of only 1 Likert item. Linear, sigmoid, 3^{rd}-degree and 5^{th}-degree polynomials are shown.



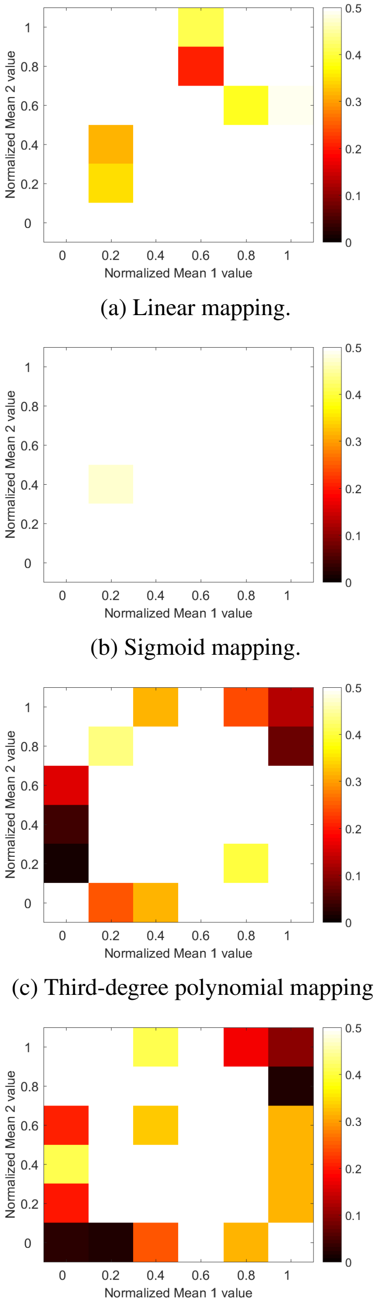
(a) Linear mapping.

(b) Sigmoid mapping.

(c) Third-degree polynomial mapping.

(d) Fifth-degree polynomial mapping.

Figure 7: These figures depict a heat map of the p-value of a Chi-squared test for independence between the number of correct and incorrect responses of the Mann-Whitney U test and t-test for experiments with 5 subjects in each of two factor levels and a Likert scale of only one Likert item. Linear, sigmoid, 3^{rd}-degree and 5^{th}-degree polynomials are shown.



(a) Linear mapping.

(b) Sigmoid mapping.

(c) Third-degree polynomial mapping.

(d) Fifth-degree polynomial mapping.

Figure 8: These figures depict a heat map of the p-value of a Chi-squared test for independence between the number of correct and incorrect responses of the Mann-Whitney U test and t-test for experiments with 30 subjects in each of two factor levels and a Likert scale of 10 Likert items. Linear, sigmoid, 3^{rd}-degree and 5^{th}-degree polynomials are shown.

family error rate. Further, if authors do use the F-test for Likert items and scales, one must state your assumptions with proof those assumptions are reasonable given prior work.

Recommendations

We provide two recommendations to the HRI community:

Authors - Applying a t-test to a one-item Likert scale may be statistically safe, but justify your assumptions with appropriate references. Further, be warned: Testing multiple, individual Likert items increases your false positive error rate. Instead, use a Likert scale or an appropriate multi-test correction.

Reviewers - Be slow to reject a paper that uses a t-test on Likert-response data. A t-test appears to be quite robust with a low false positive rate when applied to even a single Likert item. However, make sure the authors appropriately justified their analysis and controlled for the family error rate.

Limitations and Future Work

In this work, we make a number of assumptions. We assume there exists a monotonically non-decreasing mapping from Likert item responses to $\mathbb{R}$. Without such a mapping, our analysis would not be possible. The four mappings we considered, further assume that they are anti-symmetric about the point of neutral response. We also assume that each virtual subject in our Monte Carlo simulation shared a common mapping; however, this may not be true in reality. Further, we consider only four such possible mappings from a uncountably infinite set of mappings. Enumerating all such mappings is impossible. Our approach was to consider a representative set of mappings that capture a variety of behaviors. In future work, we would relax the assumption of anti-symmetry about the neutral response point. Further, we propose a set of user studies to solicit mappings subjects feel best represent their interpretation of Likert items.

An area of concern we were not able to address in this paper is the high rate of pair-wise hypothesis testing without controlling for the family error rate. In many instances, researchers would conduct an experiment with more than two factor levels and with multiple factors. While these researchers would often use an ANOVA to establish that a significance, they would then apply pair-wise F-tests or t-tests without controlling for the family error rate. Each pair-wise comparison increases the chance of a false positive. The fact that researchers commonly do not use a Scheffé test, Tukey test, or t-test applied using the Bonferonni method is alarming. Furthermore, virtually all of the papers in HRI'15 did not test the residuals to verify that the assumption of normality when conducting the ANOVA is valid. In future work, we plan to quantify the cost of these statistical practices.

Baxter et al. recently called into question the practice of null hypothesis significance testing citing poor replicability. Baxter et al. suggest using of descriptive statistics (e.g., confidence intervals) instead. While more drastic measures (e.g., Baxter et al.) are critical for the advancement of science, we are aiming for a more modest – and hopefully achievable – goal of correcting existing practices.

Conclusion

The HRI community brings together researchers from a broad array of fields, such as psychology and computer science. Most papers published in this community rely on human-subject experimentation to acquire knowledge or validate the benefit of a new robotic technology. However, there is a lack of consensus for how to construct experiment questionnaires and perform hypothesis testing through statistical analysis for responses to those questionnaires. Specifically, there is a lack of consensus for constructing and testing Likert items and Likert scales. In this paper, we survey the proceedings of HRI'15 to discuss common practices. We then conduct an extensive computational investigation via Monte Carlo simulation to test the robustness of parametric and non-parametric statistical tests for Likert items and scales. We find that the t-test is quite robust and is a reasonable method for testing even individual Likert items. However, we provide a set of recommendations to the HRI community to advocate being open-minded yet rigorous in the defense of experimental design and statistical evaluation.

Acknowledgment

The authors would like to thank Dr. Christopher Carr for being a knowledgeable, caring, admirable instructor. This work was supported by the National Science Foundation Graduate Research Fellowship Program under grant number 238835.

References

Brown, B. M., et al. 1971. Martingale central limit theorems. *The Annals of Mathematical Statistics* 42(1):59–66.

Carifio, J., and Perla, R. J. 2007. Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *Journal of Social Sciences* 3(3):106–116.

Carifio, J. 1976. Assigning students to career exploration programs by preference. *Career Education Quarterly.*

Carifio, J. 1978. Measuring vocational preferences: Ranking versus categorical rating procedures. *Career Education Quarterly* 3(2):17–28.

Donsker, M. D. 1951. An invariance principle for certain probability limit theorems. AMS.

Erdös, P., and Kac, M. 1946. On certain limit theorems of the theory of probability. *Bulletin of the American Mathematical Society* 52(4):292–302.

Glass, G. V.; Peckham, P. D.; and 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.

Hoffman, G. 2013. Evaluating fluency in human-robot collaboration. In *Proc. HRI workshop on human-robot collaboration*, volume 381, 1–8.

Nikolaidis, S.; Ramakrishnan, R.; Gu, K.; and Shah, J. 2015. Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *Proc. HRI*, 189–196.

Scheutz, M., and Arnold, T. 2016. Are we ready for sex robots? In *Proc. HRI*, 351–358.