Four Years in Review: Statistical Practices of Likert Scales in Human-Robot Interaction Studies

Mariah L. Schrum*
mschrum3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Michael Johnson*
michael.johnson@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Muyleng Ghuy*
mghuy3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Matthew C. Gombolay
matthew.gombolay@cc.gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

As robots become more prevalent, the importance of the field of human-robot interaction (HRI) grows accordingly. As such, we should endeavor to employ the best statistical practices. Likert scales are commonly used metrics in HRI to measure perceptions and attitudes. Due to misinformation or honest mistakes, most HRI researchers do not adopt best practices when analyzing Likert data. We conduct a review of psychometric literature to determine the current standard for Likert scale design and analysis. Next, we conduct a survey of four years of the International Conference on Human-Robot Interaction (2016 through 2019) and report on incorrect statistical practices and design of Likert scales. During these years, only 3 of the 110 papers applied proper statistical testing to correctly-designed Likert scales. Our analysis suggests there are areas for meaningful improvement in the design and testing of Likert scales. Lastly, we provide recommendations to improve the accuracy of conclusions drawn from Likert data.

CCS CONCEPTS

• **General and reference** → **Surveys and overviews**; *Evaluation*; Metrics.

KEYWORDS

Metrics for HRI; Likert Scales; Statistical Practices

ACM Reference Format:
Mariah L. Schrum, Michael Johnson, Muyleng Ghuy, and Matthew C. Gombolay. 2020. Four Years in Review: Statistical Practices of Likert Scales in Human-Robot Interaction Studies. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20), March 23–26, 2020, Cambridge, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3319502.3378178

*All three authors contributed equally to this research.

HRI '20, March 23–26, 2020, Cambridge, United Kingdom

ACM ISBN 978-1-4503-6746-2/20/03…$15.00
https://doi.org/10.1145/3319502.3378178

1 INTRODUCTION

The study of human-robot interaction is the interdisciplinary examination of the relationship between humans and robots through the lenses of psychology, sociology, anthropology, engineering and computer science. This all-important intersection of fields allows us to better understand the benefits and limitations of incorporating robots into a human's environment. As robots become more prevalent in our daily lives, HRI research will become more impactful on robot design and the integration of robots into our societies. Therefore, it is critical that best scientific practices are employed when conducting HRI research.

Likert scales, a commonly employed technique in psychology and more recently in HRI, are used to determine a person's attitudes or opinions on a topic [37]. Statistical tests can then be applied to the responses to determine how an attitude changes between different treatments. Such studies provide important information for how best to design robots for optimal interaction with humans. Because of the nearly universal confusion surrounding Likert scales, improper design of Likert scales is not uncommon [25]. Furthermore, care must be taken when employing statistical techniques to analyze Likert scales and items. Because of the ordinal nature of the data, statistical techniques are often applied incorrectly, potentially resulting in an increased likelihood of false positives. Unfortunately, we find the misuse of Likert questionnaires to occur frequently enough to be worth investigating.

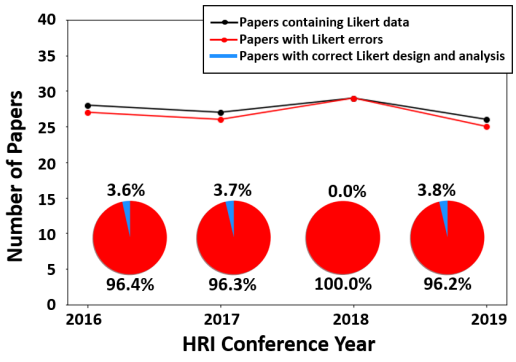

Figure 1: An overview of HRI proceedings with different types of errors when handling Likert data from 2016 - 2019.

In this paper, we 1) review the psychometric literature of Likert scales, 2) analyze the past four years of HRI papers, and 3) posit recommendations for best practices in HRI. Based upon our review of psychometric literature, we find that only 3 of 110 papers in the last four years of proceedings of HRI research properly designed and tested Likert scales. A summary of our analysis is depicted in Fig. 1. Unfortunately, this potential malpractice may suggest that the findings in 97.3% of HRI papers that based their conclusions off of Likert scales may warrant a second look.

Our first contribution is comprised of a survey of the latest psychometric literature regarding the current best practices for design and analysis of Likert scales. In cases where there is dissent or disagreement, we present both perspectives. Nonetheless, we find areas of consensus in the literature to establish recommendations for how to best design Likert scales and to analyze their data. In areas of agreement, we provide recommendations to the HRI community for how we can best construct and analyze Likert data.

Our second contribution is a survey of the proceedings of HRI 2016 through 2019 based upon the established best practices. Our review revealed that a majority of papers incorrectly design Likert scales or improperly analyze Likert data. Common mistakes are not including enough items, analyzing individual Likert items, not verifying the assumptions of the statistical test being applied, and not performing appropriate post-hoc corrections.

Our third and final contribution is a discussion of how we, as a field, can correct these practices and hold ourselves to a higher standard. Our purpose is not to dictate legalistic rules to be followed at penalty of a paper rejection. Instead, we seek to open up the floor for a constructive debate regarding how we can best establish and abide by our agreed upon best practices in our field. We hope that in doing so, HRI will continue to have a strong, positive influence on how we understand, design, and evaluate robotic systems.

Nota Bene: *We confess we have not employed best practices in our own prior work. Our goal for this paper is not to disparage the field, but instead to call out the ubiquitous misuse of a vital metric: Likert scales. We hope to improve the rigor of our own and others' statistical testing and questionnaire design so that we can stand more confidently in the inferences drawn from these data.*

2 LITERATURE REVIEW & BEST PRACTICES

Likert scales play a key role in the study of human-robot interaction. Between 2016 and 2019, Likert-type questionnaires appeared in more than 50% of all HRI papers. As such, it is imperative that we make proper use of Likert scales and are careful in our design and analysis so as not to de-legitimize our findings. We begin with a literature review to investigate the current best practices for Likert scale design and statistical testing. We acknowledge that reviews concerning the design and analysis of Likert scales have been previously conducted [11, 29, 53]. However, our analysis is the first targeted at the HRI community, and we believe it is important to ground our discussion in the current understanding of the best methods related to the construction and testing of Likert data as found in the psychometric literature.

Many of the debates surrounding Likert scale design and analysis are unsettled. As such, we present both sides of these arguments and reason through the areas of agreement and disagreement to arrive at our own recommendations for how HRI researchers can best navigate these often murky waters.

2.1 What is a Likert Scale?

Likert scales were created in 1932 by Rensis Likert and were originally designed to scientifically measure attitude [37]. A Likert scale is defined as "a set of statements (items) offered for a real or hypothetical situation under study" in which an individual must choose their level of agreement with a series of statements [31]. The original response scale for a Likert item ranged from one to five (strongly disagree to strongly agree). A seven-point scale is also common practice. An example Likert scale is shown in Fig. 2.

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
Most robots make poor teammates.	o	o	o	o	o	o	o
Most robots possess adequate decision-making capability.	o	o	o	o	o	o	o
Most robots are pleasant towards people.	o	o	o	o	o	o	o
Most robots are not precise in their actions.	o	o	o	o	o	o	o

Figure 2: This figure illustrates a portion of a balanced Likert scale measuring trust (Courtesy of [41]).

Confusion often arises around the term "scale." A Likert scale does not refer to a single prompt which can be rated on a scale from one to *n* or "strongly disagree" to "strongly agree". Rather, a Likert scale refers to a set of related prompts or "items" whose individual scores can be summed to achieve a composite score quantifying a participant's attitude toward a latent, specific topic [10]. "Response format" is the more appropriate term when describing the options ranging from "strongly disagree" to "strongly agree" [11]. This distinction is important for the following reasons. First, a high degree of measurement error arises when a participant is asked to respond only to a single prompt; however, when asked to respond to multiple prompts, this measurement error tends to average out. Second, a single item often addresses only one aspect or dimension of a particular attitude, whereas multiple items can report a more complete picture [23, 46]. Therefore, it is important to distinguish whether there are multiple items in the scale or simply multiple options in the response format. [11] emphasizes the importance of this distinction by stating that the meaning of the term scale "is so central to accurately understanding a Likert scale (and other scales and psychometric principles as well) that it serves as the bedrock and the conceptual, theoretical and empirical baseline from which to address and discuss a number of key misunderstandings, urban legends and research myths."

It is not uncommon in HRI, as well as psychometric literature, for a researcher to report that he or she employed a five-item Likert scale when in reality he or she used a single item Likert scale with five response options. To ground this distinction in an example, Fig. 2 depicts a Likert Scale with four Likert items with seven-option response format. To avoid such confusion, it is important to

be precise when describing a Likert scale as a five-option response format has a very different meaning from a five-item Likert scale. Furthermore, a set of items that prompts the user to select a rating on a bipolar scale of antonyms, i.e., human-like to machine-like, is not a true Likert scale. This is a semantic differential scale and should be referred to as such [57].

Recommendation - We recommend that HRI researchers be deliberate when describing Likert response formats and scales to avoid confusion and misinterpretation.

2.2 Design

Because HRI is a relatively new field, HRI researchers often explore novel problems for which they appropriately need to craft problem-specific scales. However, care must be taken to correctly design and assess the validity of these scales before utilizing them for research. The design of the scale is one of the least agreed upon topics pertaining to Likert questionnaires in the psychometric literature. Disagreement arises around the optimal number or response choices in an item, the ideal number of items that should comprise a scale, whether a scale should be balanced, and whether or not to include a neutral midpoint. Below, we address each topic.

Number of Response Options - Rensis Likert himself suggested a five point response format in his seminal work, *A Technique for the Measurement of Attitudes* [37]. However, Likert did not base this decision in theory and rather suggested that variations on this five-point format may be appropriate [37]. Further investigation has yet to provide a consensus on the optimal number of response options comprising a Likert item [39]. [47] found that scales with four or fewer points performed the worst in terms of reliability and that seven to nine points were the most reliable. This finding is backed up by [16] in their investigation of categorization error. [61] demonstrated via simulation that the more points a response contains, the more closely it approximates interval data and therefore recommended an 11-point response format.

This line of reasoning may lead one to believe that one should dramatically increase the number of response points to more accurately measure a construct. However, just because the data may more closely approximate interval data does not mean increasing the number of response points monotonically increases the ability to measure a subject's attitude. A larger number of response options may require a higher mental effort by the participant, thus reducing the quality of the response [5, 35]. For example, [5] conducted a study that suggested that response quality decreased above eleven response options. [52] also investigated the optimal number of response options and found that no further psychometric advantages were obtained once the number of response options rose above six and [35] suggested based on study results that the optimal number is between four and six.

Recommendation - As a general rule-of-thumb, we recommend the number of response options be between five and nine due to the declining gains with more than ten and lack of precision with less than five. However, if the study involves a large cognitive load or lengthy surveys, the researcher may want to err on the side of fewer response items to mitigate participant fatigue [47].

Neutral Midpoint - Another point of contention which influences the response number of a scale is whether or not to include a neutral midpoint. Likert, with his five-point scale, included a neutral, "undecided" option for participants who did not wish to take a positive or negative stance [37]. Some argue that a neutral midpoint provides more accurate data because it is entirely possible that a participant may not have a positive or negative opinion about the construct in question. Studies have shown that including a neutral option can improve reliability in other, similar scales [15, 26, 31, 38]. Furthermore, the lack of a neutral option precludes the participant from voicing an indifferent opinion, thus forcing him or her to pick a side which he or she does not agree with.

On the other hand, a neutral midpoint may result in users "satisficing" (i.e., choosing the option that may not be the most accurate to avoid extra cognitive strain resulting in an over-representation at the midpoint) [33]. [30] argue that "…the midpoint should be offered on obscure topics, where many respondents will have no basis for choice, but omitted on controversial topics, where social desirability is uppermost in respondents' minds."

Recommendation - We adopt the recommendation of [30], which suggests that HRI researchers utilize their best judgement based on the context of use when deciding the merits of including a neutral option in their response format. For example, if the authors are conducting a pre-trust survey to gauge a baseline level of trust before the participant has interacted with the robot, they may want to include a neutral option since some participants, especially those unfamiliar with robots, may not truly have a good sense of their own trust in robots. A neutral option would allow participants to present this sentiment. However, if a survey is being utilized to assess trust after a participant has interacted with a robot, the researchers may want to remove the neutral option, arguing that participants should have developed a sense of either trust or distrust after the interaction. Nonetheless, there may be cases when "neutral" truly is appropriate, which is why we argue in favor of researcher discretion [30].

Number of Items - The next point of contention we address is the ideal number of Likert items in a scale. In his original formulation, Likert stated that multiple questions were imperative to capture the various dimensions of a multi-faceted attitude. Based on Likert's formulation, the individual scores are to be summed to achieve a composite score that provides a more reliable and complete representation of a subject's attitude [23, 46].

Yet, in practice it is not uncommon for a single item to be used in HRI research due to the efficiency that such a short scale provides. Research into the appropriateness of single item scales has been extensively studied in marketing and psychometric literature [36]. For example, [36] investigated the use of a single-item scale for measuring a construct concluding that a single-item scale is only sufficient for simple, uni-dimensional, unambiguous objects.

Multi-item scales on the other hand are "suitable for measuring latent characteristics with many facets." [49] proposed a procedure for developing scales for evaluating marketing constructs and suggested that if the object of interest is concrete and singular, such as how much an individual likes a specific product, then a single item is sufficient. However, if the construct is more abstract and complex, such as measuring the trust an individual has for robots, then a multi-item scale is warranted. This line of reasoning is supported by [6, 17, 19]. As to the exact number of items, [19] demonstrated via simulation that at least four items are necessary for evaluation

of internal consistency of the scale. However, as suggested by [60], one should be cautious of including too many items as a large scale may result in higher refusal rates.

Recommendation - Due to the complexity of attributes most often measured in HRI (e.g., trust, sociability, usability, etc.), we recommend that researchers in the HRI community utilize multi-item scales with at least four items. The total number of items again is left to the discretion of the researcher and may depend on the time constraints and the workload that the participant is already facing. Because an average person takes two to three seconds to answer a Likert item and individuals are more likely to make mistakes or "satisfy" after several minutes, we recommend surveys not be longer than 40 items [63]. Recall that this recommendation for the number of "Likert Items" is distinct from our recommendation regarding the number of "response options," which we recommend generally be between five and nine options, as noted previously.

Scale Balance - The last aspect of scale design which we will discuss is that of balance. The question of whether the items within a scale should be balanced, i.e. there should be a parity of positive and negative statements, is one less often addressed in literature. It is believed that balancing the questionnaire can help to negate acquiescence bias, which is the phenomenon in which participants have a stronger tendency to agree with a statement presented to them by a researcher. Likert [37] advocated that scales should consist of both positive and negative statements. Many textbooks, such as [42], also state that scales should be balanced. Perhaps the most compelling evidence that balance is an important factor when developing Likert scales is provided by [51]. The authors in [51] conducted a study in which they asked participants to respond to a positively worded question to which 60% of participants agreed. They asked the same question but rephrased in a negative way and again, 60% of participants agreed. This study reveals the extent to which acquiescence bias can sway participants to answer in a particular way that is not always representative of their true feelings.

One would find this evidence to be sufficiently compelling to recommend scale balance; however, this debate is not so easily settled. Recent work suggests that although including both positively and negatively worded items reduces the effects of acquiescence bias, it may have a negative impact on the construct validity (i.e., if the scale adequately measures the construct of interest) of the scale [48, 62]. This result may be due to the fact that a negatively worded item is not a true opposite of a positively worded item. Therefore, reversing the scores of the negatively worded items and summing may have an impact on the dimensionality of the scale due to the confusion that reversed items cause [28, 56].

Recommendation - Because of a lack of consensus and the problems arising from both approaches, we do not provide a concrete recommendation to researchers about scale balance.

Validity and Reliability of Likert Prompts - Likert's original work states that the prompts of a Likert scale should all be related to a specific attitude (e.g., sociability) and should be designed to measure each aspect of the construct. Each item should be written in clear, concise language and should measure only one idea [37, 45]. This formulation helps to ensure the reliability (i.e., the scale gives repeatable results for the same participant) and the validity (i.e., the scale measures what is intended) of the scale.

A poorly formed scale may result in data that does not assess the intended hypothesis. Thus, before a statistical test is applied to a Likert scale, it is best practice to test the quality of the scale. Cronbach's alpha is one method by which to measure the internal consistency of a scale (i.e., how closely related a set of items are). A Cronbach's alpha of 0.7 is typically considered an acceptable level for inter-item reliability [54]. If the items contains few response options or the data is skewed, another method such as ordinal alpha should be employed [21].

While Cronbach's alpha is an important metric, a full item factor analysis (IFA) can be conducted to better understand the dimensionality of a scale. A scale consisting of unrelated prompts may achieve a high Cronbach's alpha for other underlying reasons or simply because Cronbach's alpha can increase as the number of items in the scale increases [24, 55]. Furthermore, a scale can show internal consistency, but this does not mean it is uni-dimensional. On the other hand, a factor analysis is a statistical method to test whether a set of items measure the same attribute and whether or not the scale is uni-dimensional. Factor analysis thus provides a more robust metric to assess the scale quality [2].

Recommendation - Due to the complex nature of scale design, we recommend that researchers utilize well-established and verified scales provided in literature when possible. Many common constructs measured in HRI can be measured with already validated scales such as the "Trust Perception Scale" for human-robot trust or the RoSAS scale for perceived sociability [12, 50]. This practice will reduce the prevalence of employing poorly designed scales. Otherwise, a thorough analysis of the internal consistency and dimensionality of new scales should be conducted when being employed to answer research questions. For in-depth instructions on how best to construct Likert scales from the ground up, please see [4, 27].

2.3 Statistical Tests

Once a scale is designed and its validity statistically verified, it is important that correct statistical tests are applied to the response data obtained from the scale. Another fiercely debated topic is whether data derived from single Likert items can be analyzed with parametric tests. We want to be clear that this controversy is not over the data type produced by Likert items but whether parametric tests can be applied to ordinal data.

Ordinal versus Interval - Previous work has demonstrated that a single Likert item is an example of ordinal data and that the response numbers are generally not perceived as being equidistant by respondents [34]. Because the numbers of a scale for Likert items represent ordered categories but are not necessarily spaced at equivalent intervals, there is not a notion of distance between descriptors on a Likert response format [14]. For example, the difference between "agree" and "strongly agree" is not necessarily equivalent to the difference between "disagree" and "strongly disagree." Thus, a Likert item does not produce interval data [7]. While it has been speculated that a large-enough response scale can approximate interval data, Likert response scales rarely contain more than 11 response points [1, 61].

Recommendation - Because a Likert item represents ordinal data, parametric descriptive statistics, such as mean and standard deviation,

are not the most appropriate metric when applied to individual Likert items. Mode, median, range, and skewness are better to report.

Parametric versus Non-Parametric - The question now becomes, given the ordinal nature of individual Likert items, is it appropriate to apply parametric tests to such data? A famous study by [22] showed that the F test is very robust to violation of data type assumptions and that single items can be analyzed with a parametric test if there is a sufficient number of response points. [34] demonstrates through simulation that ANOVA is appropriate when the single-item Likert data is symmetric but that Kruskall-Wallis should be used for skewed Likert item data. [16] also found that skew in the data results in unacceptably high errors when the data is assumed to be interval. [40] compared the use of the t-test versus the Wilcoxon signed rank test on Likert items and found that the t-test resulted in a higher Type I error rate for small sample sizes between 5 and 15. [44] made a similar comparison and also found that Wilcoxon rank-sum outperformed the t-test in terms of Type I error rates. As demonstrated by these studies, the field has yet to reach a clear consensus on whether parametric tests are appropriate, and if so when, for single Likert item data.

Likert scale data (i.e., data derived from summing Likert items) can be analyzed via parametric tests with more confidence. [22] showed that the F test can be used to analyze full Likert scale data without any significant, negative impact to Type I or Type II error rates as long as the assumption of equivalence of variance holds. Furthermore, [58] showed that Likert scale data is both interval and linear. Therefore, parametric tests, such as analysis of variance (ANOVA) or t-test, can be used in this situation as long as the appropriate assumptions hold.

Recommendation - Because studies are inconclusive as to whether parametric tests are appropriate for ordinal data, we recommend that researchers err on the conservative side and utilize non-parametric tests when analyzing Likert data. However, we also recommend that HRI researchers avoid performing statistical analysis on single Likert items altogether. As [11] so eloquently states, "one item a scale doth not make." A single item is unlikely to be the best measure for the complex constructs that are of interest in HRI research as discussed in Section 2.2. Therefore is best to avoid the ordinal vs. interval controversy altogether and instead perform analysis on a multi-item scale since Likert scales can be safely analyzed with parametric tests. If a researcher does choose to analyze an individual item, he or she should clearly state they are doing so and acknowledge possible implications. At the very least, it is recommended to test for skewness.

Post-hoc Corrections - The importance of performing proper post-hoc corrections and testing for assumptions are broadly applicable concerns, not specific to Likert data. Nevertheless, they are important considerations when analyzing Likert data and are often incorrectly applied in HRI papers.

As the number of statistical tests conducted on a set of data increases, the chances of randomly finding statistical significance increases accordingly even if there is no true significance in the data. Therefore, when a statistical test is applied to multiple dependent variables that test for the same hypothesis, a post-hoc correction should be applied. Such a scenario arises frequently when a statistical analysis is applied to individual items in a Likert scale [11]. In 2006, [3] conducted a study investigating whether individuals born under a certain astrological sign were more likely to be hospitalized for a certain diagnosis. The authors tested for over 200 diseases and found that Leos had a statistically higher probability of being hospitalized for gastrointestinal hemorrhage and Sagittarians had a statistically higher probability of a fractured humerus. This study demonstrated the heightened risk of Type I error that occurs when no post-hoc correction is applied.

There is controversy as to which post-hoc correction is best. [32] suggests applying the Bonferonni correction when only several comparisons are performed, i.e., ten or less. The authors recommend employing a different correction such as Tukey or Scheffé with more than ten comparisons to avoid the increased risk of Type II errors that stems from the conservative nature of the Bonferonni correction. [43] suggests that researchers should, instead of performing post-hoc correction, focus on reporting effect size and confidence intervals, such as Pearson's r.

Recommendation - Because of the danger that comes with performing many statistical tests without predefined comparisons we recommend that researchers always perform the proper post-hoc corrections. Due to the increased risk of Type II error that some post-hoc tests pose, we encourage researchers to also report the effect size and confidence interval to provide a more informative and holistic view of the results. In general, we recommend against pair-wise comparisons performed on individual Likert items for reasons already discussed.

Test Assumptions - Most statistical tests require certain assumptions to be met. For example, an ANOVA assumes that the residuals are normally distributed (normality) and the variances of the residuals are equal (homoscedasticity) [59]. Tests to ensure these conditions are met include the Shapiro-Wilk test for normality and Levene's test for homoscedasticity [13]. [22] argues that even when assumptions of parametric tests are violated, in certain situations, the test can still be safely applied. However, [8] counters [22] and contends that [22] failed to take into account the power of parametric tests under various population shapes and that these results should not be trusted.

Recommendation - To navigate this controversy, we suggest that researchers err on the conservative side and always test for the assumptions of the test to reduce the risk of Type I errors. If the data violates the assumptions, and the researchers decide to utilize the test despite this, they should report the assumptions of the test that have not been met and the level to which the assumptions are violated.

3 REVIEW OF HRI PAPERS

3.1 Procedures and Limitations

We reviewed HRI full papers from years 2016 to 2019, excluding alt.HRI and Late Breaking Reports, and investigated the correct usage of Likert data over these years. We considered all papers that include the word "Likert" as well as papers that employ Likert techniques but refer to the scale by a different name. We utilized the following keywords when conducting our review: "Likert", "Likert-like," "questionnaire," "rating," "scale," and "survey." After filtering based on these keywords, we reviewed a total of 110 papers. Below we report on the following categories: 1) misnomers and misleading terminology 2) improper design of Likert scales and 3) improper application of statistical tests to Likert data.

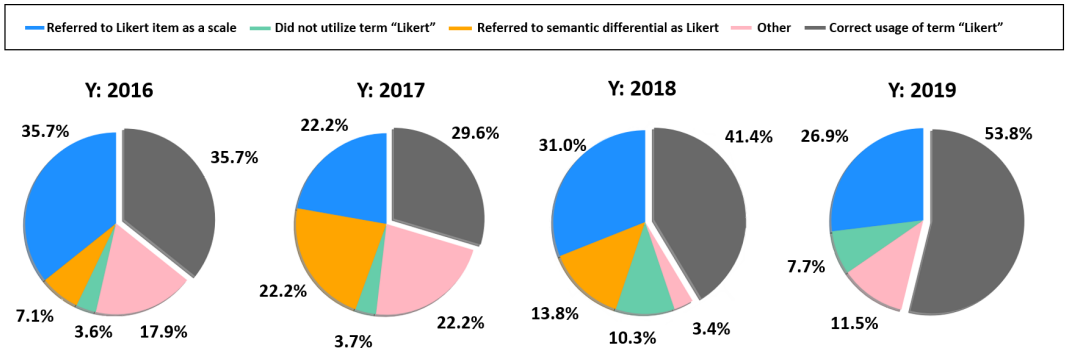

Figure 3: Common misnomer of the term "Likert Scale" within HRI Proceedings. Note: one paper in 2018 referred to a Likert item as a Likert Scale and a semantic differential scale as a Likert scale, which we counted only under the former category.

We report on the aggregate number of papers that improperly utilized the term Likert as well as papers that improperly designed Likert scales. Our observations also include papers that apply parametric tests to individual Likert items as well as papers that apply parametric tests to Likert scales but do not properly check for the assumptions of the test. Furthermore, we investigate the percentage of papers that perform statistical tests to individual items that are measuring different aspects of the same attribute but do not apply appropriate post-hoc corrections. Lastly, we report the percentage of papers that calculate the mean and standard deviation associated with individual Likert items. Fig. 1 shows the number of papers that utilized Likert-related techniques over the years under consideration. To test if the number of papers using Likert questionnaires was correlated with the year of the proceedings, we employed a Pearson correlation coefficient test, which failed to reject the null hypothesis ($t(2) = -0.617$, $p = 0.600$) that the two factors are uncorrelated. The test's assumption regarding normality was satisfied under the Shapiro-Wilk test, but homoscedasticity could not be tested as there is only one data point for each level (i.e., year). We reviewed each of these papers for correct practices. Our results illustrate the extent to which Likert data and scales are misused in HRI research and demonstrate the need for better practices to be employed to ensure the validity of results.

Throughout our review, we found ourselves limited by certain papers that did not provide enough information to properly gauge whether best practices were used. We include the count of these ambiguous papers within our results under an "Other" category. Included in this category are papers that used Likert scale questionnaires to test certain subjective metrics but did not state the number of items or other properties about the scale. This lack of detail limited our ability to determine whether their use of parametric tests were correct. In our reporting, we gave the benefit of the doubt to papers that did not report enough detail to verify the fidelity of their practices. We recommend as best practice to thoroughly report the statistical procedures used to support peer review.

3.2 Likert Misnomers

First, we report on the papers that incorrectly apply the terms "Likert" or "Likert scale." We base our analysis on the definition of Likert scale discussed in Section 2.1. Fig. 3 summarizes our findings and shows the frequency and percentages of papers that utilize each misnomer.

Mislabeling a Likert Item as a Likert Scale - The phrase "Likert scale" refers specifically to a sum across a set of related Likert items, each item measuring an aspect of the same attribute. A Likert scale prompts the user to specify their level of agreement or disagreement with a set of statements (i.e., Likert items). For the term "Likert scale" to be used, the object of reference should meet these criteria. During our review, we found that references to a single Likert item as a Likert scale are ubiquitous. For example, it is common to measure an attribute of the robot by asking a participant to rate the robot according to that trait on a Likert item response scale and to refer to this single rating as a Likert scale. While such a mistake may not have an impact on the researchers' conclusion about the relevant hypothesis, it can be misleading to the reader and may imply a more robust result than what is actually achieved. Furthermore, this misnomer may imply that parametric statistical tests are appropriate when they may not be. We found that 29% of papers labeled a Likert item as a Likert scale, and another 14% did not provide enough information about their questionnaire for us to determine whether their application of the term was accurate.

Mislabeling a Semantic Differential Scale as a Likert Scale - A "semantic continuum" consists of a set of semantic differential scales similar to how a Likert scale consists of several Likert items [57]. A semantic continuum differs from a Likert scale in that it utilizes a bipolar scale of antonyms and measures how much of a quality a specific item has. For example, a Likert item may consist of the statement "The robot makes me sad," and the user is prompted to select how much he or she agrees or disagrees with the statement. On the other hand, a semantic differential scale will prompt the user to select how the robot makes them feel, ranging from sad to happy. Multiple semantic differential scales measuring the same attribute can be summed together to form a "semantic continuum." While a semantic continuum is appropriate to utilize in many contexts, it has important inherent differences from a Likert scale. As such, we should be careful to not mislabel one as the other. Semantic continuums are specifically useful for measuring the "intensity and direction of the meaning of concepts" and have their own set

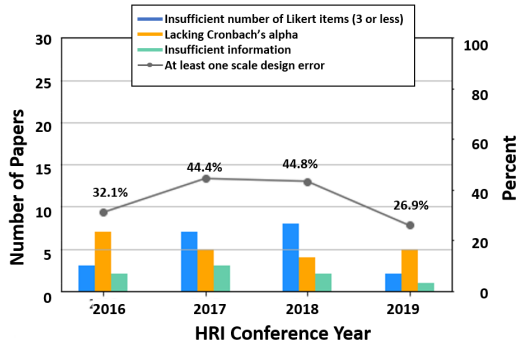

Figure 4: This figure shows the frequency of papers by year that employed improperly designed Likert scales. The percentage of papers that has at least one of these improper Likert is also reported for each year.

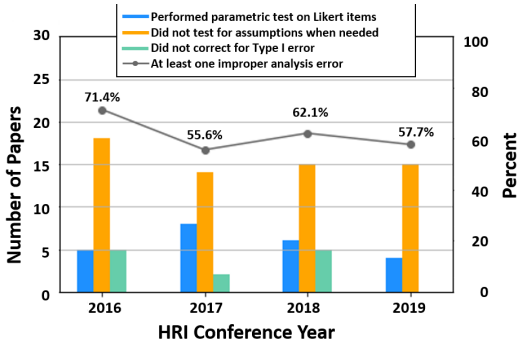

Figure 5: This figure illustrates the frequency of papers each year that incorrectly apply statistical tests on Likert data. The percentage of papers per year that incorrectly applied statistical tests is also reported.

of requirements for design as detailed in [20]. We found that an average of 7% of papers from each year adopted this misnomer.

3.3 Incorrect Design of Likert Scale

In conjunction with the improper use of the term Likert scale, we also note papers whose design or validation of a scale are questionable (see Fig. 4). Our report includes papers that utilize Likert scales with too few items, a failure to report a Cronbach's alpha, or other ambiguity within the paper's writing that could lead to disputable results. The importance of these considerations for the design of Likert scales is detailed in Section 2.2. We found that an average of 37% papers had at least one of the above errors.

3.4 Incorrect Application of Statistical Tests

In this section, we report on the recurrent ways in which statistical tests are misapplied to Likert data. We found it common for researchers to apply parametric tests to single Likert items as well as to report parametric descriptive statistics of single Likert items without stating their assumptions when doing so, both of which are not the best practice. Furthermore, papers frequently fail to check for the assumptions of parametric tests and often fail to apply appropriate post-hoc corrections. Fig. 5 summarizes our findings.

Application of Parametric Tests to Likert Items - A parametric test makes certain assumptions about the distribution from which the samples were drawn. Therefore, ANOVA, t-tests, and other parametric statistical tests are not always the most appropriate to apply to single Likert items, especially when the skew of the data is not taken into account, and their application may result in additional Type I errors. For each conference year, approximately 21% of papers with Likert data applied parametric tests when analyzing individual Likert items without testing for skewness or detailing their assumptions when doing so. Fig. 6 illustrates the number of papers that improperly analyzed single Likert items.

Inadequate Verification of Assumptions - While it is not always best practice to apply parametric tests to Likert items, it is acceptable to do so with Likert scales. This allowance is because data derived from Likert scales can be assumed to be interval in nature [18]. However, most parametric tests come with a variety of assumptions that must be met before the test can be properly applied. These assumptions test whether the data in question could have been sampled, statistically speaking, from the associated underlying distribution. For example, an ANOVA assumes that the data has been drawn from a normally distributed population, and therefore, a test for normality must be performed to verify this assumption. We observed that more than 50% of papers with Likert data from each year did not check for or report on the assumptions associated with the underlying distribution when they chose to perform a parametric test.

Inadequate Post-hoc Corrections - In general, post-hoc corrections may be performed when several dependent variables are testing the same hypotheses or when multiple statistical tests are performed on the same variables. For example, if a researcher conducts a statistical test on each individual item in a Likert scale, a correction should be applied since this is an example of testing several dependent variables that are assessing the same hypothesis. Furthermore, the chance of a Type I error increases as the number of dependent variables being tested increases. On average, we found that 11% of papers with Likert data did not account for this increased likelihood of family-wise error when they chose to perform a statistical test on individual items related to one hypothesis. For the papers that reported p-values, we performed a Bonferroni correction in order to determine the validity of the paper's result. On average 40% of the results reported in each of these papers were not significant after the adjustment. This lack of significance does not mean that the papers' conclusions are incorrect, considering the conservative nature of the Bonferonni correction. Rather, this lack suggests findings should be re-examined with proper methods.

Incorrect Reporting of Descriptive Statistics - Another common practice we found is reporting the mean and standard deviation of individual Likert items. An average of 31% of papers with Likert data from each year reported their Likert item results in this descriptive manner, most commonly through visual bar graphs. This practice is unhelpful as Likert items are ordinal data without a concept of mean or standard deviation in ordinal data. Appropriate descriptive metrics are median, mode, and range.

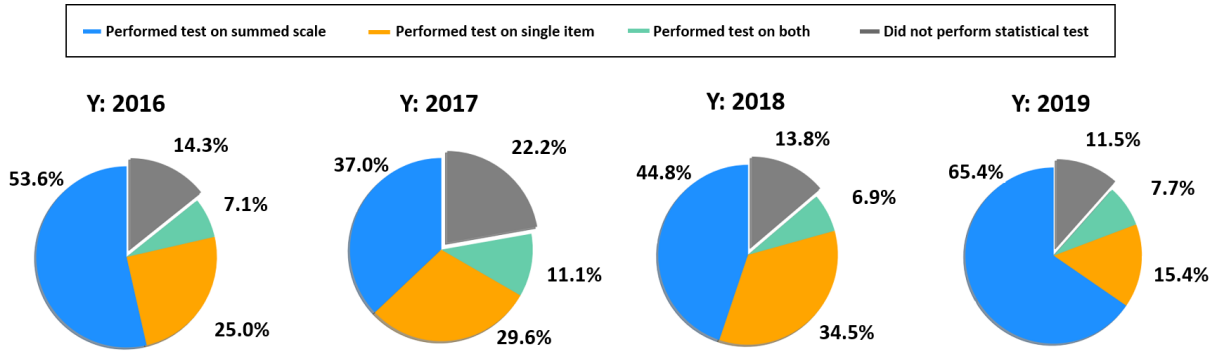

Figure 6: This figure shows the number of papers that performed statistical analysis on a Likert scale and single Likert items.

4 DISCUSSION

Our review of four years of HRI proceedings shows that nearly all relevant papers committed at least one error that could raise questions about the inferences drawn from the data. The overall trend observed between the four years does not appear to improve, leading us to believe that a call to action is warranted.

Specifically, we should seek to avoid misapplying the term Likert scale, design scales with an appropriate number of items, and test for the assumptions of the statistical analyses being applied. An in-depth review of HRI proceedings shows that the use of the term Likert scale has taken a looser connotation, as we found that roughly half of all the misnomer errors were from papers describing the response scale as a Likert scale. With respect to certain papers designing their own Likert scale for a specified metric, 18% of papers have less than four items to measure a complex construct.

Our review also shows that a large number of papers do not properly perform statistical analysis on Likert scales. Because a Likert scale is a summation across Likert items, the resulting values approximate interval data, which allows for parametric tests to be performed. However, for parametric tests to be applied, the assumption of the underlying distribution must still be tested for; and yet, 56% of papers we reviewed did not confirm this key assumption.

Finally, our analysis does not refute the conclusions of any HRI paper. Our key take-away is that we should strive for better practices so that we can be more confident in the conclusions we draw from the data. Our findings also bolster the recent support of reproducibility studies as full contributions in the field of HRI.

5 THESES

We list our recommendations to the HRI community based upon our review of the psychometric literature and in light of our findings of current HRI practices. Bold typeface is used for points made in response to the most common Likert scale issues.

- **Referring to a response scale as a Likert scale is a misnomer.** Instead, use “response format" or "response scale" when discussing the value range and reserve the term Likert scale for when referring to the entire set of items.
- Questions within a Likert scale should measure the various aspects of one and only one subjective attitude or construct.
- Likert scales should be checked for internal consistency and uni-dimensionality to ensure their reliability and validity.
- **A single Likert item should not be a sole metric for measuring a multi-faceted construct, as one statement is not generally sufficient to fully capture a complex attitude.** We recommend having at least four items.
- We encourage utilization of well-developed and validated Likert scales, e.g. RoSAS and SUS, when possible [9, 12].
- **The ordinal nature of Likert item data should be considered when selecting an appropriate statistical test.**
- It is important to systematically check for and satisfy all assumptions of the statistical tests being applied to the data.
- Experiments should be replicable: thorough detail should be provided regarding design and testing of Likert items, scales.
- **If there is more than one dependent measures supporting a single hypothesis, a correction to account for Type I error should be applied.**

6 CONCLUSION

A majority of published HRI papers rely on Likert data to gain insight into how humans perceive and interact with robots, leading Likert questionnaires to be a fundamental part of HRI studies. In this paper, we reviewed HRI proceedings from 2016-2019 and reported aggregate results of the improper use of Likert scales. Furthermore, we explored the implications of these infractions via a literature review on simulations and studies focused on incorrect design and statistical testing of Likert scales and associated data. While it is encouraging that the observed trends of the papers containing problematic usage of Likert scales and data has not increased over the last four years, it is our belief that we as a community should strive for better practices. The authors of this paper are included in this call to action. It is our hope that our recommendations are taken into consideration and that HRI researchers, authors, and reviewers employ best practices when addressing Likert data.

ACKNOWLEDGMENTS

We thank Ankit Shah for his statistical insights and support. This work was supported by institute funding at the Georgia Institute of Technology and NSF ARMS Fellowship under Grant #1545287.

REFERENCES

[1] I. Elaine Allen and Christopher A. Seaman. 2007. Likert Scales and Data Analyses.

[2] Rodrigo A. Asún, Karina Rdz-Navarro, and Jesús M. Alvarado. 2016. Developing Multidimensional Likert Scales Using Item Factor Analysis: The Case of Four-point Items. *Sociological Methods and Research* 45, 1 (2016), 109–133. https://doi.org/10.1177/0049124114566716

[3] Peter C. Austin, Muhammad M. Mamdani, David N. Juurlink, and Janet E. Hux. 2006. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *Journal of Clinical Epidemiology* 59, 9 (2006), 964–969. https://doi.org/10.1016/j.jclinepi.2006.01.012

[4] N Balasubramanian. 2012. Likert Technique of Attitude Scale Construction in Nursing Research. *Cultural Anthropology Methods* 2 (2012).

[5] A W Bendig. 1953. The Reliability of Self-Ratings as a Function of the Amount of Verbal Anchoring and of the Number of Categories on the Scale. 37, 1 (1953), 38–41.

[6] Lars Bergkvist and John R Rossiter. 2007. The Predictive Validity of Multiple-Item Versus Single-Item Measures of the Same Constructs. 2437 (2007).

[7] Phillip A Bishop and Robert L Herron. 2015. Use and Misuse of the Likert Item Responses and Other Ordinal Measures. *International journal of exercise science* 8, 3 (2015), 297–302. http://www.ncbi.nlm.nih.gov/pubmed/27182418{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4833473

[8] Clifford R Blair. 1981. A Reaction to "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analysis of Variance and Covariance". *Review of Educational Research* 51, 4 (1981), 499–507.

[9] John Brooke. 1996. SUS: a quick and dirty usability scale. In *Usability Evaluation In Industry*. CRC Press, 189–200.

[10] James Carifio and Rocco Perla. 2008. Resolving the 50-year debate around using and misusing Likert scales. (2008), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

[11] James Carifio and Rocco J. Perla. 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 (2007), 106–116. https://doi.org/10.3844/jssp.2007.106.116

[12] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The Robotic Social Attributes Scale (RoSAS): Development and Validation. *ACM/IEEE International Conference on Human-Robot Interaction* Part F1271 (2017), 254–262. https://doi.org/10.1145/2909824.3020208

[13] Flavia Chiarotti. 2004. Detecting assumption violations in mixed-model analysis of variance. *Ann Ist Super Sanità* 40, 2 (2004), 165–171.

[14] Dennis L. Clason and Thomas J. Dormody. 1994. Analyzing Data Measured By Individual Likert-Type Items. *Journal of Agricultural Education* 35, 4 (1994), 31–35. https://doi.org/10.5032/jae.1994.04031

[15] Bradley Courtenay and Craig Weidemann. 1985. The Effects of a "Don't Know" Response on Palmore's Facts on Aging Quizzes. *The Gerontologist* 2, 2 (1985), 117–181.

[16] James C. Creech and David Richard Johnson. 2019. Ordinal Measures in Multiple Indicator Models : A Simulation Study of Categorization Error Author (s): David Richard Johnson and James C . Creech Source : American Sociological Review , Vol . 48 , No . 3 (Jun ., 1983), pp . 398-407 Published by : Amer. 48, 3 (2019), 398–407.

[17] A. de Boer and P. van Lanschot, J., Stalmeier. 2004. Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? *Quality of Life Research* 13, 2 (2004), 311–320.

[18] Ben Derrick and Paul White. 2017. Comparing two samples from an individual Likert question. *International Journal of Mathematics and Statistics* (2017).

[19] Adamantios Diamantopoulos, Marko Sarstedt, Christoph Fuchs, Petra Wilczynski, and Sebastian Kaiser. 2012. Guidelines for choosing between multi-item and single-item scales for construct measurement : a predictive validity perspective. (2012), 434–449. https://doi.org/10.1007/s11747-011-0300-3

[20] Oddgeir Friborg, Monica Martinussen, and Jan H. Rosenvinge. 2006. Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences* 40, 5 (2006), 873–884. https://doi.org/10.1016/j.paid.2005.08.015

[21] Anne M Gadermann, Martin Guhn, Bruno D Zumbo, and British Columbia. 2012. Estimating ordinal reliability for Likert-type and ordinal item response data : A conceptual , empirical , and practical guide. 17, 3 (2012).

[22] Gene V Glass, Percy D Peckham, and James R Sanders. 1972. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. (1972).

[23] Joseph A. Gliem and Rosemary R. Gliem. 2003. Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales. *Midwest Research to Practice Conference in Adult, Continuing, and Community Education Calculating*, (2003). https://doi.org/10.1016/B978-0-444-88933-1.50023-4

[24] Chelsea Goforth. 2016. Using and Interpreting Cronbach's Alpha. https://data.library.virginia.edu/using-and-interpreting-cronbachs-alpha/

[25] Matthew Gombolay and Ankit Shah. 2016. Appraisal of Statistical Practices in HRI vis-á-vis the T-Test for Likert Items/Scales. In *2016 AAAI Fall Symposium Series*.

[26] Rebecca F. Guy and Melissa Norvell. 1997. The Neutral Point on a Likert Scale. *The Journal of Psychology* 95, 2 (1997).

[27] W. Penn Handwerker. 1996. Constructing Likert Scales: Testing the Validity and Reliability of Single Measures of Multidimensional Variables. *Cultural Anthropology Methods* 8 (1996).

[28] Patrick M Horan, Christine Distefano, and Robert W Motl. 2009. Wording Effects in Self-Esteem Scales : Methodological Artifact or Response Style ? 5511 (2009). https://doi.org/10.1207/S15328007SEM1003

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

[30] Robert Johns. 2006. One Size Doesn ' t Fit All : Selecting Response Scales For Attitude Items. 7289 (2006). https://doi.org/10.1080/13689880500178849

[31] Ankur Joshi, Saket Kale, Satish Chandel, and D. Pal. 2015. Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology* 7, 4 (2015), 396–403. https://doi.org/10.9734/bjast/2015/14975

[32] Hae-Young Kim. 2015. Statistical notes for clinical researchers: post-hoc multiple comparisons . *Restorative Dentistry & Endodontics* 40, 2 (2015), 172. https://doi.org/10.5395/rde.2015.40.2.172

[33] W. R. Krosnick, J. A., Narayan, S. S., & Smith. 1996. *Satisficing in surveys: Initial evidence*. San Francisco:.

[34] Bjorn Lantz. 2013. Equidistance of Likert-Type Scales and Validation of Inferential Methods Using Experiments and Simulations. 11, 1 (2013), 16–28.

[35] Jihyun Lee and Insu Paek. 2014. In Search of the Optimal Number of Response Categories in a Rating Scale. 1 (2014). https://doi.org/10.1177/0734282914522200

[36] Shing On Leung and Meng Lin Xu. 2013. Single-Item Measures for Subjective Academic Performance , Self-Esteem , and Socioeconomic Status. 8376 (2013). https://doi.org/10.1080/01488376.2013.794757

[37] Rensis Likert. 1932. A TECHNIQUE FOR THE MEASUREMENT OF ATTITUDES. *Archives of Psychology* (1932).

[38] Theodore M. Madden and Frederick J. Klopfer. 1978. The "Cannot Decide" Option in Thurstone-Type Attitude Scales. *Educational and Psychological Measurement* (1978), 259–264.

[39] Michael S. Matell and Jacob Jacoby. 1971. Is there an optimal number of alternatives for likert scale items? study 1: Reliability and validity. *Educational and Psychological Measurement* 31, 3 (1971), 657–674. https://doi.org/10.1177/001316447103100307

[40] Gary E. Meek, Ceyhun Ozgur, and Kenneth Dunning. 2007. Comparison of the t vs. Wilcoxon Signed-Rank test for likert scale data and small samples. *Journal of Modern Applied Statistical Methods* 6, 1 (2007), 91–106. https://doi.org/10.22237/jmasm/1177992540

[41] Ranjeev Mittu, Donald Sofge, Alan Wagner, and W. F. Lawless. 2016. *Robust intelligence and trust in autonomous systems*. 1–270 pages. https://doi.org/10.1007/978-1-4899-7668-0

[42] Pam Moule. 2015. *Making Sense of Research in Nursing, Health and Social Care*. SAGE Publications Ltd.

[43] Shinichi Nakagawa. 2004. A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology* 15, 6 (2004), 1044–1045. https://doi.org/10.1093/beheco/arh107

[44] Michael J Nanna. 1998. Analysis of Likert Scale Data in Disability and Medical Rehabilitation Research. 3, 1 (1998), 55–67.

[45] Tomoko Nemoto and David Beglar. 2013. Developing Likert-Scale Questionnaires. *JALT2013 Conference Proceedings* (2013).

[46] J. C. Nunnally and I. H Bernstein. 1994. *Psychometric Theory* (3rd ed.). McGraw-Hil, New York, New York, USA.

[47] Carolyn C Preston and Andrew M Colman. 2000. Optimal number of response categories in rating scales : reliability , validity , discriminating power , and respondent preferences. 104 (2000), 1–15.

[48] Lena C Quilty, Jonathan M Oakman, Evan Risko, Lena C Quilty, Jonathan M Oakman, and Evan Risko. 2009. Correlates of the Rosenberg Self-Esteem Scale Method Effects. 5511 (2009). https://doi.org/10.1207/s15328007sem1301

[49] John R Rossiter. 2002. The C-OAR-SE procedure for scale development in marketing. 19 (2002), 305–335.

[50] Kristin E. Schaefer. 2016. *Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI"*. Springer US, Boston, MA, 191–218. https://doi.org/10.1007/978-1-4899-7668-0_10

[51] Howard Schuman and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys*. Academic Press, New York, New York, USA.

[52] Leonard J Simms, Kerry Zelazny, Trevor F Williams, and Lee Bernstein. 2019. Does the Number of Response Options Matter ? Psychometric Perspectives Using Personality Questionnaire Data. 31, 4 (2019), 557–566.

[53] Basu Prasad Subedi. 2016. Using Likert Type Data in Social Science Research: Confusion, Issues and Challenges. *International Journal of Contemporary Applied Sciences* 3, 2 (2016), 2308–1365. www.ijcas.net

[54] Keith S. Taber. 2018. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education* 48,

6 (2018), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

[55] Mohsen Tavakol and Reg Dennick. 2011. Making sense of Cronbach's alpha. *International journal of medical education* 2 (2011), 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

[56] Eric van Sonderen, Robbert Sanderman, and James C. Coyne. 2013. Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain. *PLoS ONE* 8, 7 (2013), 1–7. https://doi.org/10.1371/journal.pone.0068967

[57] Tibert Verhagen, Bart van den Hooff, and Selmar Meents. 2015. Toward a better use of the semantic differential in IS research: An integrative framework of suggested action. *Journal of the Association of Information Systems* 16, 2 (2015), 108–143.

[58] Andrew J Vickers. 2019. Comparison of an Ordinal and a Continuous Outcome Measure of Muscle Soreness. 4, 1999 (2019), 709–716.

[59] Rebecca Warner. 2012. *Applied Statistics From Bivariate Through Multivariate Techniquese.* Sage Publications. 1–40 pages.

[60] Fern Willits, Gene Theodori, and A.E. Luloff. 2016. Another look at likert scales * fern k. willits. 31, August 2015 (2016), 126–139.

[61] Huiping Wu and Shing-on Leung. 2017. Can Likert Scales be Treated as Interval Scales?— A Simulation Study. *Journal of Social Service Research* 43, 4 (2017), 527–532. https://doi.org/10.1080/01488376.2017.1329775

[62] J Yamaguchi. 1997. Positive versus Negative Wording. *Rasch Measurement Transactions* 11 (1997).

[63] Ting Yan and Roger Tourangeau. 2008. Fast Times and Easy Questions : The Effects of Age , Experience and Question Complexity on Web Survey Response Times. 68, February 2007 (2008), 51–68. https://doi.org/10.1002/acp