Concerning Trends in Likert Scale Usage in Human-Robot Interaction: Towards Improving Best Practices

MARIAH L. SCHRUM*, Georgia Institute of Technology,
MUYLENG GHUY*, Georgia Institute of Technology,
ERIN HEDLUND-BOTTI, Georgia Institute of Technology,
MANISHA NATARAJAN, Georgia Institute of Technology,
MICHAEL J. JOHNSON, Georgia Institute of Technology,
MATTHEW C. GOMBOLAY, Georgia Institute of Technology,

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 in HRI research. Likert scales are commonly used metrics in HRI to measure perceptions and attitudes. Due to misinformation or honest mistakes, many 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 five years of the International Conference on Human-Robot Interaction (HRIc) (2016 through 2020) and report on incorrect statistical practices and design of Likert scales [1–3, 5, 7]. During these years, only 4 of the 144 papers applied proper statistical testing to correctly-designed Likert scales. We additionally conduct a survey of best practices across several venues and provide a comparative analysis to determine how Likert practices differ across the field of Human-Robot Interaction. We find that a venue's impact score negatively correlates with number of Likert related errors and acceptance rate, and total number of papers accepted per venue positively correlates with the number of errors. We also find statistically significant differences between venues for the frequency of misnomer and design errors. Our analysis suggests there are areas for meaningful improvement in the design and testing of Likert scales. Based on our findings, we provide guidelines and a tutorial for researchers for developing and analyzing Likert scales and associated data. We also detail a list of recommendations to improve the accuracy of conclusions drawn from Likert data.

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

Additional Key Words and Phrases: Metrics for HRI; Likert Scales; Statistical Practices

1 INTRODUCTION

The study of human-robot interaction (HRI) 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

*Both authors contributed equally to this research.

Authors' addresses: Mariah L. Schrum, mschrum3@gatech.edu, Georgia Institute of Technology, 266 Ferst Dr NW, Atlanta, Georgia, 30332, ; Muyleng Ghuy, mghuy3@gatech.edu, Georgia Institute of Technology, 266 Ferst Dr NW, Atlanta, Georgia, 30332, ; Erin Hedlund-Botti, erin.botti@gatech.edu, Georgia Institute of Technology, 266 Ferst Dr NW, Atlanta, Georgia, 30332, ; Manisha Natarajan, mnatarajan30@gatech.edu, Georgia Institute of Technology, 266 Ferst Dr NW, Atlanta, Georgia, 30332, ; Michael J. Johnson, michael.johnson@gatech.edu, Georgia Institute of Technology, 85 5th St NW, Atlanta, Georgia, 30308, ; Matthew C. Gombolay, matthew.gombolay@cc.gatech.edu, Georgia Institute of Technology, 266 Ferst Dr NW, Atlanta, Georgia, 30332,

2573-9522/2022/11-ART $15.00
https://doi.org/10.1145/3572784

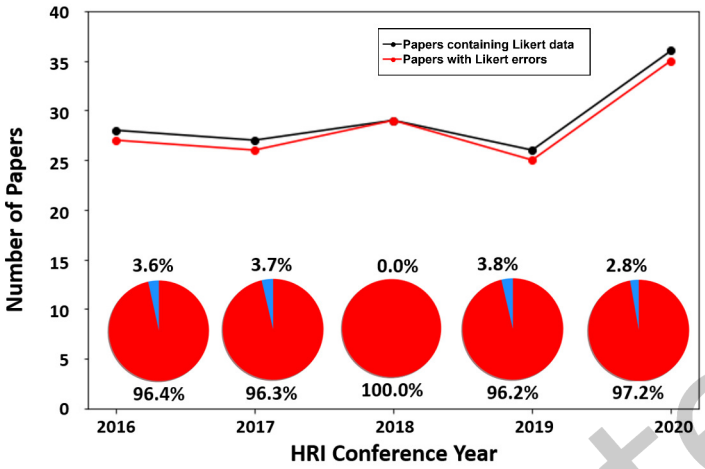

Fig. 1. The line graph compares the total number of papers with Likert data to the number of papers with Likert errors in the HRIc proceedings from 2016 - 2020. The pie charts compare the percentage of papers with and without errors for each year.

benefits and limitations of incorporating robots into a human's environment. As robots become more prevalent in our daily lives, HRI research will have a greater impact 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 measure a person's attitudes or opinions on a topic [80]. 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 [46]. 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 in the field of HRI to be worth investigating.

In this paper, we 1) review the psychometric literature of Likert scales, 2) analyze best practices for the past five years of papers in the International Conference on Human-Robot Interaction (HRIc)* proceedings, 3) investigate best practices across venues, 4) provide a tutorial and 5) posit recommendations for best practices in HRI. We extend the work of [110] by providing a more in depth review of psychometric literature with regards to scale validity and reliability. We also add a tutorial to make our recommendations more accessible to researchers. We additionally provide a full review of the 2020 International Conference on Human-Robot Interaction [7] and a more thorough analysis of all five years as well as an analysis of practices across venues in the field of HRI. Based upon our review of best practices from psychometric literature, we find that only 4 of 144 papers in the last five years of HRIc proceedings properly designed and tested Likert scales and that less than 2% of papers across four HRI venues in 2019 and 2020 employed best practices. A summary of our analysis for the HRIc proceedings is depicted in Figure 1 and a summary of our

*In this paper we distinguish between the acronyms for the field of HRI and the Conference on HRI (HRIc) with a lowercase c.

analysis across venues is depicted in Figure 9. Unfortunately, this deviation from best practices may suggest that the findings in more than 98% of HRI papers that based their conclusions off of Likert scales may warrant a second look. Our intent is to highlight the widespread past misuse of Likert scales in the field to motivate better practices in the future. We hope that the tutorial and best practices detailed in this paper will provide researchers in the field with clarity and resources for the correct usage of Likert scales.

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 are able to find many areas of consensus in the literature to establish recommendations for how to best design Likert scales and to analyze their data. In these areas of agreement, we provide recommendations to the HRI community for how to best construct and analyze Likert data.

Our second contribution is a survey of the proceedings of HRIc 2016 through 2020 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 contribution is an analysis and comparison of best practices across four venues in the field of HRI for the years 2019 and 2020. Our investigation suggests that improper practices are prevalent throughout the field of HRI and that use of best practices positively correlates with impact score and negatively correlates with acceptance rate.

Our fourth contribution is a tutorial for HRI researchers to reference when designing and analyzing Likert scales and associated data. This guide provides a list of steps which researchers should comply with when designing a Likert scale to ensure reliability and validity of the scale. We also include a comprehensive list of validated Likert scales from prior literature for commonly measured attitudes in HRI. Lastly, we provide a flowchart for researchers to follow when analyzing Likert data to ensure that best practices are followed.

Our fifth 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 always 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 this data.*

2 LITERATURE REVIEW & BEST PRACTICES

Likert scales play a key role in the study of human-robot interaction. Between 2016 and 2020, Likert-type questionnaires appeared in more than 50% of all HRIc papers. As such, it is imperative that we, as members of the HRI community, make proper use of Likert scales and are careful in our design and analysis so as not to de-legitimize our collective 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 [26, 58, 113]. However, our analysis is the first targeting the HRI community, and we

	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

Fig. 2. This figure illustrates a portion of a balanced Likert scale measuring trust (Courtesy of [86]).

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 [80]. 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 [71]. 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 Figure 2.

Response Format - 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 [25]. "Response format" is the more appropriate term when describing the options ranging from "strongly disagree" to "strongly agree" [26]. 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 random measurement error tends to average out. We note that multiple items will reduce random error, but not necessarily systematic error. Second, a single item often addresses only one aspect or dimension of a particular attitude, whereas multiple items can report a more complete picture [44, 94]. Therefore, it is important to distinguish whether there are multiple items in the scale or simply multiple options in the response format. [26] 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 incorrectly refer to a response format as a Likert scale. To ground this distinction in an example, Figure 2 depicts a Likert scale with four Likert items and a 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

Distinguishing Between Other Metrics - A psychometric tool should only be labeled as a Likert scale if it complies with the description in this section. Various scales exist that are similar to Likert scales but differ in important ways. For example, a "semantic continuum" consists of a set of semantic differential scales similar to how a Likert scale consists of several Likert items [118]. 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 object 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 they agree or disagree 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 (for further reading on the differences in data arising from semantic continuums versus Likert scales, please see [41]). For example, semantic continuums are specifically useful for measuring the "intensity and direction of the meaning of concepts" and have their own set of requirements for design as detailed in [41]. As such, we should be careful to not mislabel one as the other. Additionally, scales such as NASA TLX and SWAT that utilize different or additional methods for calculating composite scores should be distinguished from standard Likert scales via terms such as "Likert variant" or "Likert-like" [52, 103].

> *Recommendation - We recommend that HRI researchers be deliberate when describing Likert response formats and scales to avoid confusion and misinterpretation and to only refer to scales that meet the criteria in Section 2.1 as Likert Scales.*

2.2 Design and Development

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 of 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. The development of the scale also requires rigorous validity and reliability analysis. 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* [80]. However, Likert did not base this decision in theory and rather suggested that variations on this five-point format may be appropriate [80]. Further investigation has yet to provide a consensus on the optimal number of response options comprising a Likert item [82]. [98] 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 [70] in their investigation of categorization error. [125] 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 [17, 77]. For example, [17] conducted a study that suggested that response quality decreased above eleven response options. [112] 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 [77] 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 [98].*

Response Format Label - By the formal definition, a Likert scale response format should be labeled from "strongly disagree" to "strongly agree" [80]. Although there is little evidence in the psychometric literature to suggest that this choice of label is superior to other choices, other response format labels have not been widely studied and therefore are not as well understood. Furthermore, a review conducted by [35] suggests that the response format label may have an impact on data quality and interpretation.

There is further debate about the label of the midpoint (see below for a discussion about inclusion versus omission of a midpoint). Likert's original scale utilized the label "undecided" for the midpoint [80]. However, researchers have suggested that either "neutral" or "neither agree nor disagree" are better alternative to "undecided" as "undecided" may represent an absence of opinion and therefore not comply with the ordinal nature of the response format [30].

Prior work [49] has also investigated the labeling of Smiley-o-Meter scales which are Likert-like scales commonly employed in research with children (see Section 5 for more detail). The standard Smiley-o-Meter utilizes smiley faces as labels, typically ranging from sad to happy. Hall and Hume conducted several studies with various response labels and found that children rarely selected the negative ratings, perhaps because children are tuned to more positive attitudes [49]. To solve this issue, the researchers created the Five Degree of Happiness scale which utilizes varying degrees of happy faces for the response labels which produced higher quality responses by encouraging the use of all scale points in studies with children.

> *Recommendation - We recommend that authors adhere to the "strongly disagree" to "strongly agree" response format label when possible, as this has been thoroughly validated. Further, we recommend that authors utilize either "neutral" or "neither agree nor disagree" when labeling a midpoint to maintain the ordinal nature of the scale. When deviating from this label, we recommend that authors instead refer to their scale as "Likert-like" to differentiate it from the classical Likert scale. When soliciting responses from children, utilize the Five Degrees of Happiness scale [49]*

Neutral Midpoint - Another point of contention which influences the response format 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 [80]. Some argue that the inclusion of 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 [32, 48, 71, 81]. Furthermore, the lack of a neutral option precludes the participant from voicing an indifferent opinion, thus forcing them to pick a side which they 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) [75]. The authors in [69] 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 [69], 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, based on the notion 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 [69].*

Overall Response Format Design - The number of response options and the response format labels are intrinsically linked. The number of response options inevitably influences the choice of response labels. The more response options, the more difficult it is to assign a label to each option. Typically scales with many response options must rely on anchor labels with either number labels or no labels for intermediate options. Prior work has investigated the differences that arise in fully labeled versus partially labeled scales as well as the effect of gradation of (dis)agreement (e.g., a five-point scale has two gradations whereas a seven-point has three) when labeling the response scale [122]. Weijters et al. found that a fully labeled scale led to higher quality responses [122]. Thus, the authors recommend in situations of opinion measurement and scale development to utilize either a five-point or seven-point *fully labeled* response format. These findings are supported in other studies which demonstrate that a fully labeled scale produces higher reliability [12, 74].

> *Recommendation - In alignment with our above recommendations on number of response options and response format labels and the recommendations provided in [122], we recommend that authors utilize a five-point or seven-point fully labeled response format to achieve high-quality responses. In a five-point response format, authors should label the options "strongly disagree," "disagree," either "neutral" or "neither agree nor disagree," "agree," and "strongly agree." In a seven-point response format, authors should label the options "strongly disagree", "disagree", "slightly disagree," either "neutral" or "neither agree nor disagree," "strongly agree," "agree," and "slightly agree." However, we recognize that little research has been conducted on the exact nature of the response format labels and therefore, we provide this recommendation only as a soft guideline.*

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 [44, 94].

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 [79]. For example, [79]

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." [105] 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 [18, 34, 38]. As to the exact number of items, [38] demonstrated via simulation that at least four items are necessary for evaluation of internal consistency of the scale. However, as suggested by [123], one should be cautious of including too many items, as a large scale may result in higher refusal rates (i.e., more unanswered questions).

> *Recommendation - Due to the complexity of attributes most often measured in the field of 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 [128]. 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 [80] advocated that scales should consist of both positive and negative statements. Many textbooks, such as [87], 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 [111]. The authors in [111] 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 [100, 127]. 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 [57, 117].

> *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 Scales - 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 are both contingent on the guidelines listed above. For example, [44] found that a single item scale decreased reliability, and as discussed by [33], using scales with five to seven response options improves reliability and validity. Additionally, 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 [80, 90]. This formulation helps to ensure the reliability and the validity of the scale. Therefore, to improve validity and reliability, researchers should closely adhere to the above recommendations when designing Likert scales.

Even if these guidelines are followed, ensuring the reliability and validity of a scale is not a simple task. Rigorous analysis and repeated studies should be conducted to confirm the legitimacy of the scale before use. When designing a scale, an initial pool of items (two times to five times the size of the desired size of the final scale) should be created [22]. Items should be derived from theory and prior work. Content validity of each item should be verified by experts in the field. Items can then be eliminated via factor analysis and measures of internal validity to form the final scale. Common methods for item reduction include the Classical Test Theory (CTT) and Item Response Theory (IRT) which rely on item difficulty index, discrimination index, inter-item and item-total correlations, and distractor efficiency analysis to determine the best items in the pool [22, 50, 73, 124]. If, after CTT or IRT has been applied to the scale, the number of items are less than the recommended minimum of four items, the researchers should create additional items based on theory and expert knowledge.

> *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 the field of HRI can be measured with already validated scales such as the "HRI Trust Scale" for human-robot trust [126] or the RoSAS scale for perceived sociability [27]. This practice will reduce the prevalence of employing poorly designed scales. A thorough list of verified scales for common HRI metrics can be found in Section 5.*

Internal Consistency and Dimensionality - 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 [115]. If the items contains few response options or the data is skewed, another method, such as ordinal alpha, should be employed [42]. Cronbach's alpha alone does not ensure the reliability of a scale. For example, 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 [45, 116]. Therefore, it is also good practice to utilize a test-retest method in which the scale is tested within the same population across multiple points in time in addition to reporting Cronbach's alpha [102]. Furthermore, recent work has suggested that other internal consistency metrics such as McDonald's omega coefficient, ω, may provide better estimates of reliability [36, 101]. For further discussion on this topic, please reference Deng and Chan [36].

While Cronbach's alpha and other reliability tests are important metrics, a full item factor analysis (IFA) should be conducted to better understand the dimensionality of a scale. 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 [13].

Additionally, factor analysis is crucial in scale development to determine which items load on each factor. A factor, in this context, describes a latent variable. For example, in the RoSAS scale, a tool commonly used in HRI research, these latent variables are warmth, competence, and discomfort [27]. During scale development, factor eigenvalues, derived from Factor Analysis (FA), are utilized to determine the importance of each factor. Factors with an eigenvalue greater than one are retained. Factor loading values are commonly employed to determine which items load onto each factor. It is recommended to retain items that have a factor loading of above 0.4 because these items explain more than 10% of the variance in the data [22, 108].

Recommendation - If researchers choose to create their own scales rather than employing well-established scales from prior work, a thorough analysis of the internal consistency and dimensionality of new scales should be conducted before deployment. Factoring loading values for individual items should be at least 0.4 and factors with eigenvalues greater than one should be retained. For in-depth instructions on how best to construct Likert scales from the ground up, please see [22, 51, 114]. Please also see Section 5 for further reference.

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 [76]. 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 [31]. 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 [20]. 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 [10, 125].

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 [43] 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 are a sufficient number of response points. [76] demonstrates through simulation that ANOVA is appropriate when the single-item Likert data is symmetric but that Kruskal-Wallis should be used for skewed Likert item data. [70] also found that skew in the data results in unacceptably high errors when the data is assumed to be interval. [83] 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. [89] 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. [43] 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, [119] 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 on full Likert scales 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 single Likert items. However, we also recommend that HRI researchers avoid performing statistical analysis on single Likert items altogether. As [26] 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 appropriate assumptions are met. If a researcher does choose to analyze an individual item, they 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 applies to all data and is 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 [78]. 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 [26]. In 2006, [14] 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. [72] 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. The authors of [88] suggest 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) [121]. Tests to ensure these conditions are met include the Shapiro-Wilk test for normality and Levene's test for homoscedasticity [28]. [43] argues that even when assumptions of parametric tests are violated, in certain situations, the test can still be safely applied. However, [21] counters [43] and contends that [43] 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 HRIC PAPERS

3.1 Procedures and Limitations

We reviewed HRIc full papers from years 2016 to 2020, 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." We then omitted papers that did not utilize Likert or Likert-like scales. For example, we omitted papers that used the word "scale" in a context unrelated to a questionnaire (e.g., size or weight measurement) and papers that utilized questionnaires that are a different form from Likert or Likert-like. After filtering based on these keywords and exclusion criteria, we reviewed a total of 144 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.

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. Figure 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 ($t(3) = 1.2942$, $p = 0.2862$) 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

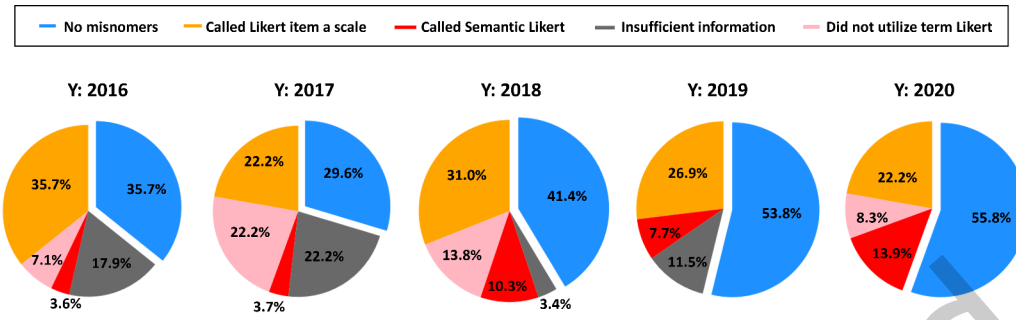

Fig. 3. Each pie chart shows the misuse of the term "Likert Scale" within the HRIc Proceedings for a year in the range 2016 - 2020. Note: one paper in 2018 referred to a Likert item as a Likert Scale and a semantic differential scale as a Likert scale, which we counted only under the former category.

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 and reproducibility.

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. Figure 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 28% of papers labeled a Likert item as a Likert scale, and another 11% did not provide enough information about their questionnaire for us to determine whether their application of the term was accurate.

Mislabeling a Semantic Continuum as a Likert Scale - Semantic continuums, while closely related to Likert scales are not equivalent to Likert scales. For example, 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 [118]. Therefore, a distinction, which is often overlooked, should be made when employing these two tools. We found that an average of 8% 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 not in keeping with best practices (see Figure 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 43% papers had at least one of the above errors.

Too few items - In Figure 5, we show the total number of scales with one, two, three and four plus items from the five years of HRIc papers which we surveyed. The majority of scales (61%) have an improper number of items (i.e., fewer than four) to capture complex attributes. Scales for which not enough information was reported to determine the number items were not included in these results. Scales that fail to include at least four items may not be capable of accurately measuring complex attributes as discussed in Section 2.2.

Improper response format label - While there is little evidence that straying from the formal definition of a Likert response format label ("strongly disagree" to "strongly agree") affects the validity of the results, we encourage authors to refer to Likert scales that utilize variants such as "low" to "high" or "not at all" to "very much" as Likert-like scales to prevent confusion. In our review, we found that 30.6% of papers employed alternate response format labels.

Failure to report Cronbach's alpha - Cronbach's alpha is an important measure of internal consistency and should be reported for every scale employed in a study. We found in the five years of HRIc papers that we evaluated, 24% of papers failed to report Cronbach's alpha. When reporting these results, we did not include papers that only utilized one-item scales as it is not possible to report Cronbach's alpha for a single item. Reporting Cronbach's alpha provides reviewers and readers with an estimate of the internal consistency and thus the reliability of the scale within the context of the sample population; therefore, authors should ensure they report this metric.

Failure to employ verified scales - As discussed in Section 2.2, Likert scales should undergo a rigorous verification process before being employed to answer research questions. While many of the scales utilized by the papers we reviewed complied with our guidelines for number of items, number of response formats, reporting Cronbach's alpha etc., a significant number of scales did not undergo the verification process to ensure reliability and validity. In total, we find that 72.5% of the scales were not verified via the methods discussed in Section 2.2 to ensure both validity and reliability. 9.8% of the scales are verified in previous work but are altered in some way (i.e., items are removed). 17.7% of scales utilized in the papers we reviewed were verified for reliability and validity.

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. Figure 6 summarizes our findings.

Incorrect 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

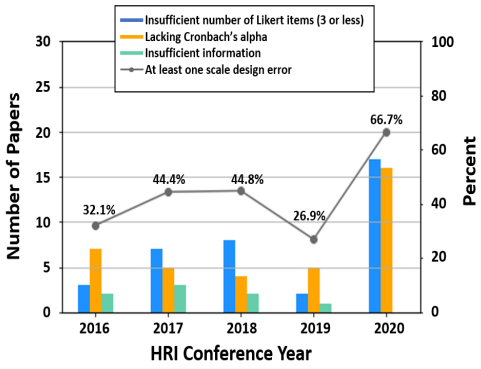

Fig. 4. The line graph represents the percentage of papers by year in the HRI Proceedings that employed improperly designed Likert scales. The bars display the frequency of each type of scale error. We observed a large increase in the percent of papers that committed at least one error in the year 2020 for HRIc.

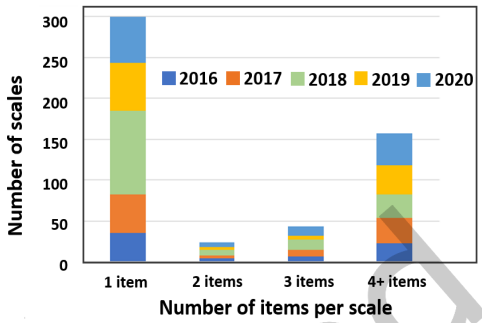

Fig. 5. Each bar, broken down by year, represents the total number of scales with one, two, three, and four or more items. Scales with less than four items are not capable of capturing complex attributes as discussed in Section 2.2. The number of items comprising each scale was reported for 524 total scales.

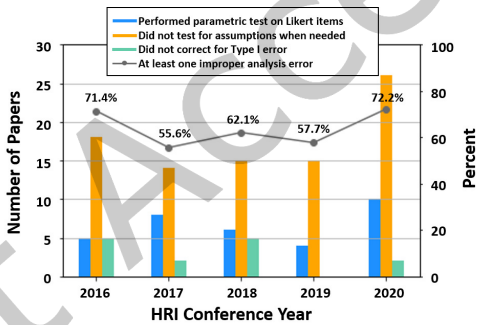

Fig. 6. The line graph reports the percentage of papers per year that incorrectly applied statistical tests in the HRI Proceedings. The bars illustrate the frequency of papers that made each type of statistical error on Likert data.

result in additional Type I errors. For each conference year, approximately 22% of papers with Likert data applied parametric tests when analyzing individual Likert items without testing for skewness or detailing their assumptions when doing so. Figure 7 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 [37]. 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

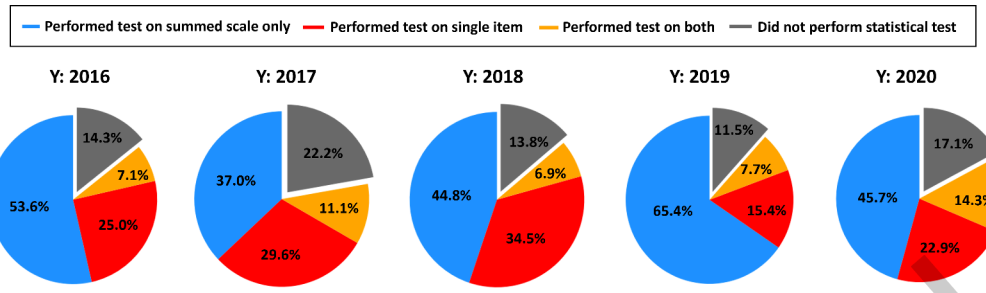

Fig. 7. Each pie chart shows the percent of papers that performed statistical analysis on a Likert scale and single Likert items for the HRIc Proceedings for years 2016 to 2020.

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 that equals the number of items 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 10% 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.

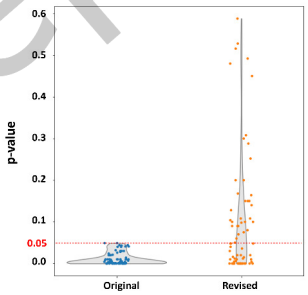

Fig. 8. The density plots of the original p-values reported in the papers (left) and the revised p-values after appropriate post-hoc correction has been applied for all venues (right) show that many fewer p-values were significant after post-hoc correction.

For the papers that reported p-values and failed to conduct proper post-hoc corrections, we performed a Bonferroni correction in order to investigate the validity of the paper's result. The Bonferonni correction is defined as $\frac{\alpha}{m}$ where m is the number of hypotheses being tested and α is the significance level [53]. When determining significance, this is equivalent to maintaining the significance level and multiplying the p-value by m. Therefore, we corrected the p-value based on the number of hypotheses tested for each instance of improper post-hoc analysis. In Figure 8, we show a density plot of the original and corrected p-values. On average 46% 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 of significance 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 29% of papers with Likert data from each year reported their Likert item results in this descriptive manner, most commonly through visual bar graphs displaying the means and standard deviations of Likert score.

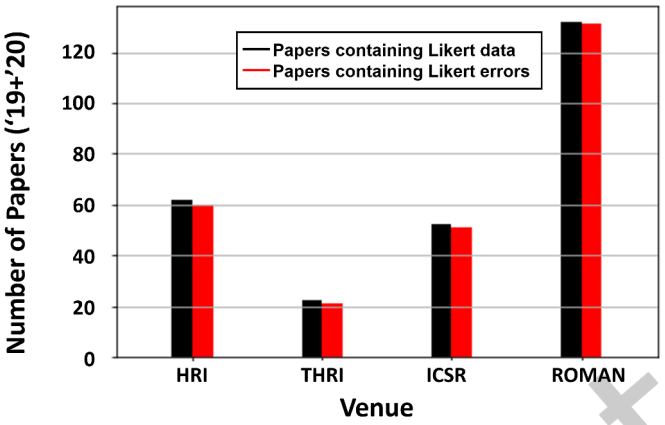

Fig. 9. This figure compares the overall errors for each venue. Each venue has a high percentage of Likert errors.

This practice is unhelpful as Likert items are ordinal data without a concept of mean or standard deviation. Appropriate descriptive metrics are median, mode, and range.

4 COMPARISON OF LIKERT PRACTICES ACROSS VENUES

We next review papers and report results from the following four venues in the field of human-robot interaction to determine if Likert practices differ across four venues: 1) *Proceedings of the International Conference on Human-Robot Interaction* (HRIc)) [1–3, 5, 7], 2) *Transactions on Human-Robot Interaction* (THRI) [59–66], 3) *Proceedings of the International Conference on Social Robotics* (ICSR) [107, 120], and 4) *Proceedings of the International Conference on Robot and Human Interactive Communication* (RO-MAN) [4, 6] for the years 2019 and 2020. We utilize the same criteria detailed in Section 2.3 to conduct our review. In the following sections we report on 1) misnomers and misleading terminology, 2) improper design of Likert scales, and 3) improper application of statistical tests to Likert data. We compare the prevalence of each of these errors across venues via a Chi-squared analysis, and we investigate whether the use of best practices is related to the type of venue (i.e., journal or conference), impact score, acceptance rate, and total number of accepted papers. In doing so, we seek to determine if the frequency of errors per venue has an effect on how often papers are cited from that venue as measured via impact score, which is defined as the yearly average number of citations divided by number of published articles. Additionally, we investigate if the selectivity of a venue (as measured via acceptance rate) or volume of papers accepted (as measured via total number of papers accepted), has an impact on the frequency of Likert-related errors.

Figure 9 shows the total number of papers accepted for each venue and the number of papers that employ correct practices. Each venue only employed best practices in less than 2% of all papers containing Likert data across the years 2019 and 2020. To determine which metrics impact error rate, we conduct a correlation analysis between the venue's impact score, acceptance rate, and total number of accepted papers at each venue and the percent of papers that employ incorrect practices. We apply Shapiro Wilk's test for normality and non-constant variance score test for homoscedasticity to ensure that our data meets parametric assumptions. Because all data passed

Table 1. This table lists the correlation coefficients and confidence intervals between the metrics of interest and frequency of errors. We do not include THRI in our analysis for acceptance rate as this number is not reported.

	Misnomer	Design	Analysis
Impact Score	$r = -0.55$ [-0.99, 0.87]	$r = -0.97$ [-0.99, -0.20]	$r = -0.37$ [-0.98, 0.92]
Acceptance Rate	$r = 0.76$ [-0.12, 0.97]	$r = 0.83$ [0.05, 0.98]	$r = -0.09$ [-0.84, 0.78]
Number Accepted	$r = 0.31$ [-0.47, 0.85]	$r = 0.61$ [-0.17, 0.92]	$r = -0.14$ [-0.77, 0.63]

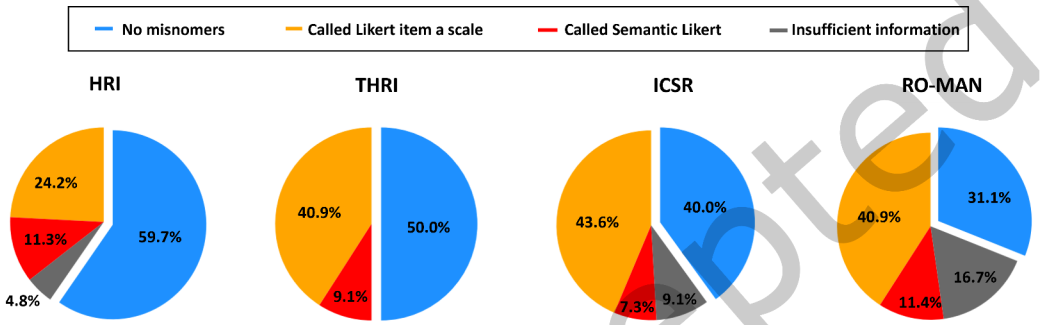

Fig. 10. Each pie chart highlights the type of misuse of the term "Likert Scale" for each venue.

these tests, we employ Pearson's correlation for all correlation analyses. As THRI does not report acceptance rate, we exclude it from our acceptance rate correlation analysis. We note that our sample size is relatively small (four venues) and therefore these results are only exploratory in nature, and, due to the small sample size, we do not report p-values. Instead we focus on the general trends that arise from these metrics. Our correlation results are reported in Table 1 and we provide a more in depth discussion of these results in the following sections.

4.1 Likert Misnomers

Figure 10 shows the percent of papers across venues that incorrectly employed the name Likert, broken down into the categories detailed in Section 2.3. We find that HRIc and THRI correctly use the term Likert $\geq 50\%$ of the time whereas ICSR and RO-MAN do so $\leq 40\%$ of the time. Across all venues, the most common mistake was referring to a Likert item as a scale. We note that all papers in THRI provided sufficient information with regards to their use of the term Likert whereas at least 4% of papers in the conference venues did not provide suffient information. We hypothesize that this result is due to the more rigorous peer review process employed by journals. Because there are several rounds of reviews that occur in the journal acceptance process, reviewers have ample opportunity to ensure that adequate information is provided by the authors. Additionally, unlike HRIc, ICSR, and RO-MAN, THRI does not impose a page limit, thereby providing more space for authors to provide the necessary information about their scales.

Furthermore, we find that the percent of papers that incorrectly utilize the name Likert negatively correlates with impact score ($r = -.55$), suggesting that venues that improperly employ the term Likert are less likely to have their papers cited. We hypothesize that this result is due to the fact that papers which incorrectly employ the term Likert are more confusing and difficult to understand than those which properly employ the term Likert, resulting in a lower rate of citations. Alternatively, harder to understand papers may degrade the reputation of the conference, resulting

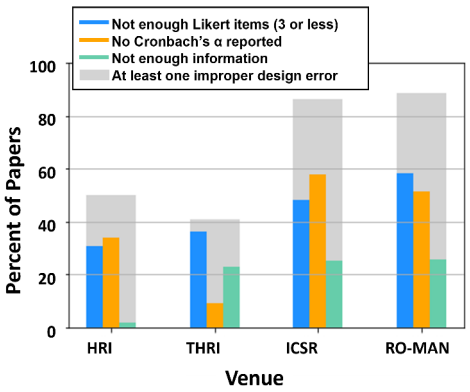

Fig. 11. For each venue, the percentage of papers with a scale design error is represented by the wide grey bar. The type of error is further broken down and represented by the thinner colored bars.

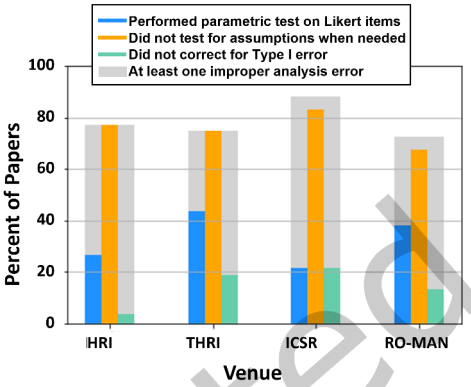

Fig. 12. The wide grey bars represent the percent of papers with analysis errors across the four venues, and the thinner colored bars represent each type of error. These percentages are calculated as the number of papers with an error divided by the total number of Likert papers that performed statistical analysis on Likert data.

in less awareness of the papers in the conference. We also find that improper use of the term Likert positively correlates with both acceptance rate ($r = .76$) and total number of accepted papers ($r = .31$), indicating that venues with higher acceptance rates and higher overall number of accepted papers may be more likely to accept low quality work with a higher frequency of misnomer errors.

Next, we conduct a Chi-squared analysis to determine if the frequency of Likert misnomers significantly differs across venues. We apply Yates correction when frequencies are less than five to mitigate overestimation of statistical significance. In our analysis, we find that there is a statistically significant difference in the frequency of misnomer errors among the different venues ($p = .0017$). Therefore, the data suggests that the number of misnomer errors is dependent on the venue. We employ a Chi-squared analysis with a Bonferonni correction to determine pairwise significance between venues. Based on this analysis, we find that HRIc had significantly fewer misnomer errors compared to RO-MAN ($p < .001$). We did not find significance between the other pairwise comparisons.

4.2 Incorrect Design of Likert Scale

We next analyze the proportion of papers that utilized improperly designed Likert scales as shown in Figure 11. HRIc and THRI had the lowest error rate with regards to improper design. We find that for HRIc and ICSR, the most common design error was failure to report Cronbach's α whereas the most common error in THRI and RO-MAN was not enough Likert items in a scale. Interestingly, we find that THRI did not report sufficient information about scale design 36% of the time, which is greater than HRIc (31%) while still being below ICSR (48%) and RO-MAN (58%). Unlike in the case of Likert misnomers, the more stringent peer review journal process did not seem to reduce insufficient information in THRI for scale design, suggesting that reviewers do not ensure that authors provide sufficient information with regards to scale design even with the more rigorous

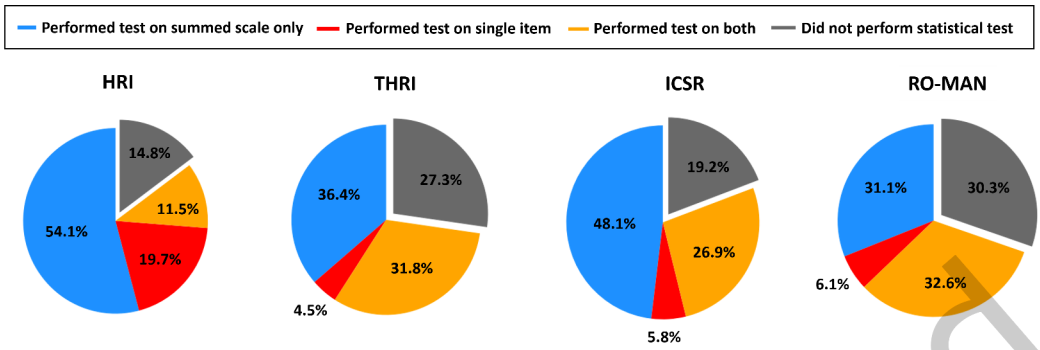

Fig. 13. Each pie chart shows the percentage of analysis error results for each venue.

journal process. This finding is particularly concerning, considering that improperly designed scales can result in an increase in Type I and Type II errors as discussed in Section 2.2.

Next, we conduct a correlation analysis between the metrics of interest and the frequency with which venues utilized improperly designed scales. We find a strong negative correlation between impact score and percentage of time that papers employ improperly designed scales ($r = -.97$). This finding may suggest that venues which place more emphasis on employing proper metrics when deciding to accept or reject a paper are viewed more favorably by other researchers and therefore are more likely to be cited. Our findings also show a strong relationship between acceptance rate and improper scale design ($r = .83$) and between number of papers accepted and improper scale design ($r = .61$). We hypothesize that this relationship is due to the fact that venues that accept more papers devote less resources to reviewing individual papers, resulting in low quality papers with poorly designed scales being accepted when a more careful review would have resulted in a rejection.

Moreover, we applied a Chi-squared test to determine if there is a significant difference between frequency of scale design errors across venues. Our analysis suggests that the venue does significantly impact the frequency of scale design errors ($p < .001$). After a pairwise comparison and Bonferonni post-hoc correction, we find that HRIc had significantly less design errors compared to both ICSR ($p < .001$) and RO-MAN ($p < .001$) and that THRI also had fewer errors compared to ICSR ($p < .001$) and RO-MAN ($p < .001$).

4.3 Incorrect Application of Statistical Tests

Lastly, we investigate incorrect application of statistical tests across venues as shown in Figures 12 and 13. We find that HRIc correctly performed tests on summed scales rather than on a single item more often than other venues. THRI performed analysis on both a summed scale and single items 31.8% of the time, which is more often than both HRIc (11.5%) and ICSR (26.9%). We hypothesize that, because authors have more space for additional analysis in a journal, they chose to perform analysis on both individual items as well as a summed scale despite this analysis being incorrect. Due to the fact that all venues had similarly high error rates, a Chi-squared analysis did not find any significant difference between the venues for frequency of incorrect application of statistical tests.

In Figure 12, we show improper analysis results broken down into additional categories. We see a fairly consistent distribution of errors across venues. Because of this fairly even distribution, we do not find strong correlations between the percent of analysis errors and impact score, acceptance

rate, or total number of accepted papers. Across all venues, a large portion of papers (74%) did not properly check for assumptions before employing a parametric test while relatively fewer papers (13%) failed to account for Type I error when making multiple comparisons or when testing the same hypothesis multiple times. Failure to account for Type I error is a particularly egregious practice due to the increased risk of reporting effects that do not exist. Therefore, we investigated this error further by performing the post-hoc corrections when possible via a Bonferonni correction to determine how the original reported p-values compared to the corrected p-values as shown in Figure 14. For the papers that did not conduct a necessary post-hoc correction in HRIc, for years 2019 and 2020, 89% of original reported p-values were no longer significant after applying the Bonferonni correction. 12.5% of the reported p-values in THRI, 55.6% in ICSR, and 26.1% in RO-MAN were no longer significant after applying the post-hoc correction. Our findings suggest that the significant results reported in these papers should be re-evaluated with the proper methods to verify the validity of the papers' conclusions.

5 LIKERT-LIKE SCALES

Many Likert-like scales (i.e., scales that share similarities with Likert scales but do not meet the criteria described in Section 2.1) are commonly used in HRI research. In this section we provide a brief overview of Likert-like scales, their proper uses, and how they have been employed in HRI research.

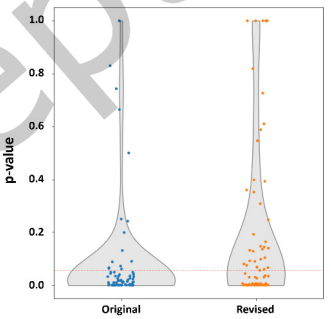

Fig. 14. The density plots of the original p-values reported in the papers (left) and the revised p-values after appropriate post-hoc correction has been applied for all venues (right) show that fewer p-values were significant after post-hoc correction.

Semantic Continuum - Originally introduced by Osgood, Suci, and Tannenbaum in 1957, the semantic differential scale is a tool employed for measuring attitudes on a bipolar continuum rather than on a scale from strongly disagree to strongly agree [96]. As discussed in Section 2.1, a semantic continuum is the term given to a scale comprised of several semantic differential scales. Likert scales and semantic continua both capture quantitative data of multi-dimensional, complex attitudes. However, there are various differences in how they are structured and perceived by responders. Several considerations that must go into designing semantic differential scales (e.g., the selection of adjectives, scale layout, and relevance to participants) are detailed in [9]. In some cases, a semantic format may be more appropriate than a Likert scale, as detailed in [41]. It is therefore imperative for researchers to be aware of the differences between semantic continua and Likert scales so as to select the most appropriate tool for a specific study. In the 2020 HRIc proceedings, 10.6% of papers properly employed semantic continuums versus only 2.2% in 2016. 10.6% of THRI papers, 13.3% of ICSR papers, and 10.0% of RO-MAN papers employed semantic continuums for years 2019 and 2020.

NASA TLX - Workload is often a useful measure in HRI when comparing various algorithm implementations. The NASA Task Load Index (NASA TLX) is a tool designed by the Human Performance Group at NASA Ames Research for measuring perceived workload [52]. Subjective workload assessments are split into six factors– mental demand, physical demand, temporal demand, performance, effort, and frustration level. The participant rates the six dimensions of workload on a sliding rating scale from very low to very high, typically measured from 0 to 100 when scored. The participants then select in pairwise comparisons which dimension is more relevant to the workload of the task being evaluated, creating a weighting for each item [52]. For the HRIc proceedings, in

2020, 4.5% of papers utilized a NASA TLX scale to measure perceived workload, 10.4% in 2019, 6% in 2018, 7.8% in 2017 and 2.2% in 2016. Across the years 2019 and 2020, 3.0% of THRI papers, 3.0% of RO-MAN papers, and <1% of ICSR papers employed NASA TLX.

Smiley-o-Meter - Smiley-o-meters also known as Smiley Face Likert Scales (SFL) are commonly employed tools in research with children. When surveying children, "the child must be able/provided with an effective method to communicate the judgment made" about their experience [16]. An SFL has proved to be an effective method for a child to communicate their attitude towards a construct. [49] provides an animated 5-point response format smiley-o-meter and shows via multiple studies that this scale results in low satisficing and sample variance in children. In our review, we find that one paper from 2020 and one paper from 2017 employ SFL in their research in HRI. No papers in THRI for years 2019 or 2020 employed smiley-o-meter scales. 1.5% of ICSR papers and <1% of RO-MAN papers utilized the scale.

6 TUTORIAL FOR DESIGNING AND ANALYZING LIKERT SCALES AND DATA

6.1 Scale Design

In this section we present a tutorial for designing and analyzing Likert scales. The ability to draw correct statistical conclusions from Likert data begins with proper scale design. We give an in-depth discussion of proper scale design in Section 2.2 based on recommendations in psychometric literature. Here we provide an overview of the steps one should take as well as important equations to employ when designing and validating a scale. Additionally, we provide illustrative guides to aid in proper scale design. Figure 15, adapted from [22], provides a step-by-step guide of the important design considerations and validations that should be performed when designing a new Likert scale. To ensure the validity of the scale, researchers should be careful to not skip any steps.

Steps for Verifying Scales - The steps for constructing scales and ensuring their validity and reliability are numerous and at times complicated. Here we outline several of the important steps and equations one should employ when verifying scales.

Item Reduction - Three criteria should be taken in to account when eliminating items from a pool: item difficulty index, item discrimination index, and inter-item correlation [22]. Item discrimination index is an important measure and defines how well an item discriminates between different individuals, scored between −1 and 1. This can be done by determining the point biserial correlation score between each item and the total score of the questionnaire. Items that fail to discriminate or discriminate negatively should be removed. The biserial correlation index is calculated according to Eq. 1 [47].

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}} \tag{1}$$

Factor Extraction - The factors of a Likert scale are typically extracted via factor analysis which is based on a regression model. This model is described in Eq. 2-3 in which X are the observed variables, μ the means, Λ the factor loadings, and F the factors. E is the error matrix, R the correlations and D the unique variance.

$$X = \mu + \Lambda F + E \tag{2}$$

$$(R - D) = AA^{'} \tag{3}$$

Tests of Dimensionality - Tests of dimensionality exist to verify that the factors extracted do not vary when taken from two independent samples or from the same sample at two different time

points. Techniques for this assessment include the Chi-squared test of exact fit and Root Mean Square Error of Approximation (RMSEA).

Tests of Reliability - Cronbach's alpha and test-retest methods are two means by which to verify the reliability of a scale. Cronbach's alpha is described in Eq. 4 in which n is the number of questions, V_i the variance of scores in each question and V_{test} is the total variance of the overall scores.

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum V_i}{V_{test}}\right) \tag{4}$$

Tests of Validity - While it is best to ensure that the designed instruments pass all tests of validity (content, criterion, face validity, etc.), evidence has shown that satisfying at least two of the different forms of construct validity is enough to ensure a valid scale [22]. For example, researchers can employ the multi-trait-multi-method matrix estimation tool for calculating the convergent and discriminant factors of constructs validity. For a detailed process of how to perform this calculation, please see [24]. In addition, researchers can employ a correlation analysis among Likert items to ensure construct validity.

Verified Scales - We as HRI researchers sympathize with the time-consuming nature of the scale development process. Therefore, we provide a list of previously-verified scales for measuring various attitudes common in HRI research that we encourage researchers to utilize rather than designing and validating their own scales. We note that changes to these scales (e.g., combining scales, removing items, etc.) would require researchers to go through the process of validating the new scale following the steps provided in Figure 15. We ensured that the scales on the following list have undergone rigorous analysis including factor analysis, establishing Cronbach's alpha and other important considerations. The list includes both Likert scales, semantic differential scales, and scales such as NASA TLX and SWAT which are widely used variants of Likert.

Trust in Robots/Technology

- The HRI Trust Scale * [126]
- Trust Perception Scale - HRI ‡ [109]
- Trust in Automated Systems* [68]
- Propensity to Trust in Technology Scale (PTT) * [67]

General Trust

- Interpersonal Trust Scale (ITS)* [106]
- Faith-in-People Scale* [104]
- Propensity to Trust scale* [40]

Anthropomorphism

- Godspeed subscale† [15]
- Anthropomorphism Tendency Scale (ATS)* [29]
- The Uncanny Valley Effect Scale‡ [55]

Usability

- System Usability Survey (SUS)* [23]

Engagement

- User Engagement Survey (UES)* [95]

Sociability

- Heerink Toolkit Questionnaire* [54]
- The Robots Social Attributes Scale (RoSAS)* [27]
- The Robot Conversation Scale† [85]

Attitude/Bias Towards Robots/Technology

- Automation-Induced Complacency Potential* [84]
- Negative Attitude Towards Robots Scale (NARS)* [92]
- Frankenstein Syndrome Questionnaire (FSQ)* [93]
- Multi-Dimensional Robot Attitude Scale * [91]
- Technology-Specific Expectations Scale (TSES)* [11]

*Likert Scale
‡Variant of Likert
†Semantic Differential Scale

- Attitude Towards Technology Scale* [39]

Likeability

- Godspeed subscale† [15]
- RoSAS warmth subscale* [27]

Fluency

- Fluency in HRI Scale* [56]

Workload

- NASA Task Load Index‡ [52]
- Subjective Workload Assessment Technique‡ [103]

Self-Efficacy

- Self-Efficacy in Human-Robot-Interaction Scale (SE-HRI)* [99]
- Multi-Dimensional Robot Attitude Scale Self-Efficacy Subscale* [91]

Acceptance of Robot/Technology

- Technology-Specific Satisfaction Scale (TSSS)* [11]
- Robot Acceptance Survey (RAS)* [19]
- Ethical Acceptability Scale* [97]

6.2 Lack of Verified Scale

If no scale exists to measure the attribute of interest, ideally authors would go through the appropriate steps to create and validate an appropriate scale (Fig. 15). However, we recognize that doing so can be prohibitively time consuming and may hinder the progress of timely research. Therefore, we suggest that if authors choose to employ a scale that has not been validated, then they should take steps to ensure that readers understand the limitations associated with this decision. Authors should discuss how they crafted the items (e.g., reference the prior work from which the items are derived). They should report metrics related to reliability and validity as discussed in Section 2.2. Furthermore, authors should ensure to state in the limitations section that their scale is exploratory and has not been verified in prior work. Efforts should be made in future work to validate the scale and reproduce study findings.

6.3 Scale Analysis

Once an appropriately designed scale is utilized to collect data, proper analysis must be conducted to ensure statistically sound results. Various decisions must be made by researchers when conducting analysis on Likert data, including whether to use parametric or non-parametric tests, which assumptions must be checked and whether or not to apply a post-hoc correction. As such, confusion often arises as to the proper method of statistical analysis. In Figure 16 we present a flowchart detailing the proper route through the maze of statistical analysis to which researchers should adhere when analyzing Likert data.

Choosing a Statistical Test: When choosing a statistical test, researchers should first determine the level of measurement of their data. If they are are applying a test to a single item (which we do not recommend), then the researchers should employ a non-parametric test. Otherwise, researchers should check that their data passes parametric assumptions. If this is the case then researchers may apply a parametric test.

Applying a Post-hoc Correction: When testing multiple hypotheses or making pairwise comparisons, researchers should ensure that they apply a post-hoc correction.

Reporting Results: We recognize that space is often limited to report full results. Therefore, we recommend that authors report the results relevant to the research questions in the main paper and thorougly report additional results (e.g., tests for assumptions) in the Appendix. We recommend that authors provide a table that details the independent variable, dependent variable, statistical test employed, and the results of the tests for assumptions. An example is provided in Table 2.

Table 2. Example of a table to include in the Appendix describing the variables, tests for assumptions, results, and statistical tests applied.

Study 1				
DV	IV	Test	Normality	Homoscedasticity
Dependent Var. #1	Independent Var. #1	Friedman's	$W = x.x$, $p = x.x$	$F(x, x) = x.x$, $p = x.x$
Dependent Var. #2	Independent Var. #2	ANOVA	$W = x.x$, $p = x.x$	$F(x, x) = x.x$, $p = x.x$

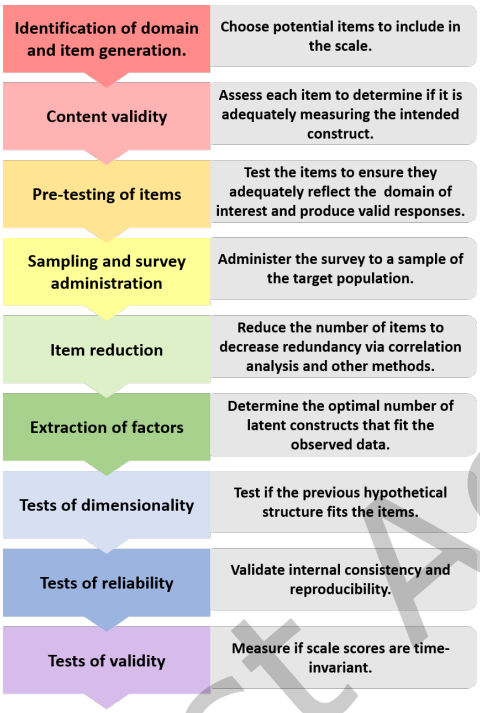

Fig. 15. Researchers should undergo the steps in the list above when designing and validating a scale [22].

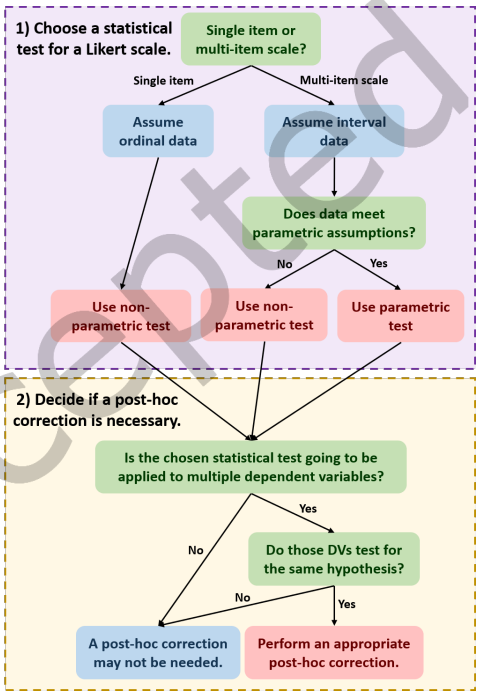

Fig. 16. This flowchart depicts the proper steps to follow when analyzing Likert data, including how to choose the appropriate statistical test and post-hoc correction.

7 DISCUSSION

In 2015, Nosek et al. [8] conducted a study in which a group of researchers replicated 100 psychology studies. Only 36% yielded significant results compared to the original 97% that found significance. These results caused many in the scientific community to question the validity and integrity of the field of psychology. Our fear is that the field of HRI may face similar criticism if we do not adhere to best practices.

Our review of five years of HRIc 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 five 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" and design scales with an appropriate number of items. 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 incorrect scale design, 25% 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 assumptions of the underlying distribution must still be tested for; and yet, over 60% of papers we reviewed did not confirm these key assumptions.

Based on our review of papers across venues, we find that Likert-related errors are prevalent across the field of HRI. Our analysis shows that less than 2% of papers in HRIc, THRI, ICSR, and RO-MAN employed correct practices in the years 2019 and 2020. We find that error rates correlate with impact score suggesting that improper practices may impact a venue's reputation. ICSR and RO-MAN frequently utilize the term Likert incorrectly with more than 60% of papers having a misnomer error. Despite the rigorous review process employed by journals, THRI failed to provide sufficient information 36% of the time with regards to scale design. Furthermore, we find that a large portion of papers did not check for assumptions (74%) and many did not account for increased risk of Type 1 errors (13%), both of which can lead to effects being reported when none exist.

Finally, we want to emphasize that our analysis does not refute and is not intended to refute the conclusions of any HRI paper. Our key takeaway 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 for reproducibility studies as full contributions in the field of HRI.

8 RECOMMENDATIONS FOR BEST PRACTICE

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.
- Items 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 [23, 27].
- **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 and scales.

- **If there is more than one dependent measure supporting a single hypothesis, a correction to account for Type I error should be applied.**

9 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 HRIc proceedings from 2016-2020 and THRI, ISCR, and RO-MAN proceedings from 2019-2020 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. Unfortunately, the number of papers that misused Likert surveys greatly increased in 2020 and we find that incorrect practices are prevalent across venues. Therefore, 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. To this end, we include a tutorial to aid HRI researchers when utilizing Likert scales and data in their research.

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] 2016. *HRI '16: The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (Christchurch, New Zealand). IEEE Press.

[2] 2017. *HRI '17: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria). ACM, New York, NY, USA.

[3] 2018. *HRI '18: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA). ACM, New York, NY, USA.

[4] 2019. *28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019, New Delhi, India, October 14-18, 2019*. IEEE. https://ieeexplore.ieee.org/xpl/conhome/8951224/proceeding

[5] 2019. *HRI '19: Proceedings of the 2019 ACM/IEEE International Conference on Human-Robot Interaction* (Daegu, South Korea). ACM, New York, NY, USA.

[6] 2020. *29th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2020, Naples, Italy, August 31 - September 4, 2020*. IEEE. https://ieeexplore.ieee.org/xpl/conhome/9219088/proceeding

[7] 2020. *HRI '20: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom). Association for Computing Machinery, New York, NY, USA.

[8] Alexander A. Aarts, Joanna E. Anderson, Christopher J. Anderson, Peter R. Attridge, Angela Attwood, Jordan Axt, Molly Babel, Štěpán Bahník, Erica Baranski, Michael Barnett-Cowan, Elizabeth Bartmess, Jennifer Beer, Raoul Bell, Heather Bentley, Leah Beyan, Grace Binion, Denny Borsboom, Annick Bosch, Frank A. Bosco, and Sara D. Bowman. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716. https://doi.org/10.1126/science.aac4716

[9] Jayne Al-Hindawe et al. 1996. Considerations when constructing a semantic differential scale. *La Trobe papers in linguistics* 9, 7 (1996), 1–9.

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

[11] Patrícia Alves-Oliveira, Tiago Ribeiro, Sofia Petisca, Eugenio Di Tullio, Francisco S. Melo, and Ana Paiva. 2015. An empathic robotic tutor for school classrooms: Considering expectation and satisfaction of children as end-users. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9388 LNCS, October (2015), 21–30. https://doi.org/10.1007/978-3-319-25554-5_3

[12] Duane F. Alwin and Jon A. Krosnick. 1991. The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes. *Sociological Methods & Research* 20, 1 (1991), 139–181. https://doi.org/10.1177/0049124191020001005 arXiv:https://doi.org/10.1177/0049124191020001005

[13] 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

[14] 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

[15] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[16] Alice Bell. 2007. Designing and testing questionnaires for children. *Journal of Research in Nursing* 12, 5 (2007), 461–469. https://doi.org/10.1177/1744987107079616 arXiv:https://doi.org/10.1177/1744987107079616

[17] 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. *Journal of Applied Psychology* 37, 1 (1953), 38–41.

[18] Lars Bergkvist and John R. Rossiter. 2007. The Predictive Validity of Multiple-Item versus Single-Item Measures of the Same Constructs. *Journal of Marketing Research* 44, 2 (2007), 175–184. https://doi.org/10.1509/jmkr.44.2.175

[19] Linda M. Beuscher, Jing Fan, Nilanjan Sarkar, Mary S. Dietrich, Paul A. Newhouse, Karen F. Miller, and Lorraine C. Mion. 2017. Socially assistive robots: Measuring older adults' perceptions. *Journal of Gerontological Nursing* 43, 12 (2017), 35–43. https://doi.org/10.3928/00989134-20170707-04

[20] 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

[21] 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.

[22] Godfred O. Boateng, Torsten B. Neilands, Edward A. Frongillo, Hugo R. Melgar-Quiñonez, and Sera L. Young. 2018. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health* 6, June (2018), 1–18. https://doi.org/10.3389/fpubh.2018.00149

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

[24] Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin* 56, 2 (1959), 81.

[25] James Carifio and Rocco Perla. 2008. Resolving the 50-year debate around using and misusing Likert scales. *Med Educ* 42, 12 (2008), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

[26] 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

[27] 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

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

[29] Matthew G. Chin, Ryan E. Yordon, Bryan R. Clark, Tatiana Ballion, Michael J. Dolezal, Randall Shumaker, and Neal Finkelstein. 2005. Developing and Anthropomorphic Tendencies Scale. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 49, 13 (2005), 1266–1268. https://doi.org/10.1177/154193120504901311 arXiv:https://doi.org/10.1177/154193120504901311

[30] Seung Youn Yonnie Chyung, Katherine Roberts, Ieva Swanson, and Andrea Hankinson. 2017. Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale. *Performance Improvement* 56 (11 2017), 15–23. Issue 10. https://doi.org/10.1002/pfi.21727

[31] 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

[32] 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.

[33] John Dawes. 2008. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research* 50, 1 (2008), 61–77. https://doi.org/10.1177/147078530805000106

[34] Angela de Boer, J.J.B. Lanschot, Peep Stalmeier, J.W. Sandick, Jan Hulscher, Hanneke (JCJM) Haes, and M.A.G. Sprangers. 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 : an international journal of quality of life aspects of treatment, care and rehabilitation* 13 (04 2004), 311–20. https://doi.org/10.1023/B:QURE.0000018499.64574.1f

[35] Anna DeCastellarnau. 2018. A classification of response scale characteristics that affect data quality: a literature review. *Quality and Quantity* 52, 4 (2018), 1523–1559. https://doi.org/10.1007/s11135-017-0533-4

[36] Lifang Deng and Wai Chan. 2017. Testing the Difference Between Reliability Coefficients Alpha and Omega. *Educational and Psychological Measurement* 77 (4 2017), 185–203. Issue 2. https://doi.org/10.1177/0013164416658325

[37] Ben Derrick and Paul White. 2017. Comparing two samples from an individual Likert question. *International Journal of Mathematics and Statistics* 18 (2017), 1–13.

[38] 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. *Journal of the Academy of Marketing Science* 40, 3 (2012), 434–449. https://doi.org/10.1007/s11747-011-0300-3

[39] Steve Edison and Gary L Geissler. 2003. Measuring attitudes towards general technology: Antecedents, hypotheses and scale development. *Journal of Targeting, Measurement and Analysis for Marketing* 12 (2003), 137–156.

[40] M. Lance Frazier, Paul D. Johnson, and Stav Fainshmidt. 2013. Development and validation of a propensity to trust scale. *Journal of Trust Research* 3, 2 (2013), 76–97. https://doi.org/10.1080/21515581.2013.820026

[41] 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

[42] 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. *Practical Assessment, Research & Evaluation* 17, 3 (2012), 1–13.

[43] 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. *Review of Educational Research* 42, 3 (1972), 237–288. https://doi.org/10.3102/00346543042003237

[44] Joseph A. Gliem and Rosemary R. Gliem. 2003. Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales. In *Midwest Research to Practice Conference in Adult, Continuing, and Community Education*. Columbus, 82–88. https://doi.org/10.1016/B978-0-444-88933-1.50023-4

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

[46] 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*.

[47] S. Das Gupta. 1960. The General Model Consider a universe Z consisting of two sub-universes Zo and Z ˜ , and. *Psychometrika* 25, 4 (1960), 393–408.

[48] Rebecca F. Guy and Melissa Norvell. 1997. The Neutral Point on a Likert Scale. *The Journal of Psychology* 95, 2 (1997), 199–204.

[49] Lynne Hall, Colette Hume, and Sarah Tazzyman. 2016. Five Degrees of happiness: Effective Smiley Face Likert scales for evaluating with children. *Proceedings of IDC 2016 - The 15th International Conference on Interaction Design and Children* (2016), 311–321. https://doi.org/10.1145/2930674.2930719

[50] Ronald Hambleton and H Swaminathan. 2013. *Item Response Theory: Principles and Applications*. Springer Science & Business Media.

[51] W. Penn Handwerker. 1996. Constructing Likert Scales: Testing the Validity and Reliability of Single Measures of Multidimensional Variables. *Cultural Anthropology Methods* 8, 1 (1996), 1–7. https://doi.org/10.1177/1525822X960080010101

[52] Sandra Hart and Lowell Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Human Mental Workload* 43, 5 (1988), 138–179. https://doi.org/10.1007/s10749-010-0111-6

[53] Winston Haynes. 2013. *Bonferroni Correction*. Springer New York, New York, NY, 154–154. https://doi.org/10.1007/978-1-4419-9863-7_1213

[54] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. 2009. Measuring acceptance of an assistive social robot: A suggested toolkit. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication* (2009), 528–533. https://doi.org/10.1109/ROMAN.2009.5326320

[55] Chin Chang Ho and Karl F. MacDorman. 2017. Measuring the Uncanny Valley Effect: Refinements to Indices for Perceived Humanness, Attractiveness, and Eeriness. *International Journal of Social Robotics* 9, 1 (2017), 129–139. https://doi.org/10.1007/s12369-016-0380-9

[56] Guy Hoffman. 2019. Evaluating Fluency in Human-Robot Collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218. https://doi.org/10.1109/THMS.2019.2904558

[57] Patrick M Horan, Christine Distefano, and Robert W Motl. 2003. Wording Effects in Self-Esteem Scales: Methodological Artifact or Response Style? *Structural Equation Modeling: A Multidisciplinary Journal* 10, 3 (2003), 435–455. https://doi.org/10.1207/S15328007SEM1003

[58] 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

[59] Odest Chadwicke Jenkins and Selma Sabanovic (Eds.). 2019. *J. Hum.-Robot Interact.* 8, 4 (2019).
[60] Odest Chadwicke Jenkins and Selma Sabanovic (Eds.). 2019. *J. Hum.-Robot Interact.* 8, 3 (2019).
[61] Odest Chadwicke Jenkins and Selma Sabanovic (Eds.). 2019. *J. Hum.-Robot Interact.* 8, 2 (2019).
[62] Odest Chadwicke Jenkins and Selma Sabanovic (Eds.). 2019. *J. Hum.-Robot Interact.* 8, 1 (2019).
[63] Odest Chadwicke Jenkins and Selma Sabanovic (Eds.). 2020. *J. Hum.-Robot Interact.* 9, 4 (2020).
[64] Odest Chadwicke Jenkins and Selma Sabanovic (Eds.). 2020. *J. Hum.-Robot Interact.* 9, 3 (2020).
[65] Odest Chadwicke Jenkins and Selma Sabanovic (Eds.). 2020. *J. Hum.-Robot Interact.* 9, 2 (2020).
[66] Odest Chadwicke Jenkins and Selma Sabanovic (Eds.). 2020. *J. Hum.-Robot Interact.* 9, 1 (2020).
[67] Sarah A. Jessup, Tamera R. Schneider, Gene M. Alarcon, Tyler J. Ryan, and August Capiola. 2019. *The Measurement of the Propensity to Trust Technology.* Vol. 11575. Springer International Publishing. 476–489 pages. https://doi.org/10.1007/978-3-030-21565-1
[68] Jiun-Yin Jian. 1998. Foundations for Empirically Determined Scale of Trust in Automated Systems.
[69] Robert Johns. 2005. One Size Doesn't Fit All: Selecting Response Scales For Attitude Items. *Journal of Elections, Public Opinion and Parties* 15, 2 (2005), 237–264. https://doi.org/10.1080/13689880500178849
[70] David Richard Johnson and James C. Creech. 1983. Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review* 48 (1983), 398.
[71] 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
[72] 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
[73] Theresa Kline. 2014. *Psychological Testing: A Practical Approach to Design and Evaluation.* SAGE Publications, Inc., Thousand Oaks, California, Chapter Classical Test Theory: Assumptions, Equations, Limitations, and Item Analyses, 91–106. https://doi.org/10.4135/9781483385693.n5
[74] Jon A. Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5, 3 (1991), 213–236. https://doi.org/10.1002/acp.2350050305 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350050305
[75] Jon A. Krosnick, Sowmya Narayan, and Wendy R. Smith. 1996. Satisficing in surveys: Initial evidence. *New Directions for Evaluation* 1996, 70 (1996), 29–44. https://doi.org/10.1002/ev.1033
[76] Bjorn Lantz. 2013. Equidistance of Likert-Type Scales and Validation of Inferential Methods Using Experiments and Simulations. *Electronic Journal of Business Research Methods* 11 (2013), 16–28.
[77] Jihyun Lee and Insu Paek. 2014. In Search of the Optimal Number of Response Categories in a Rating Scale. *Journal of Psychoeducational Assessment* 32, 7 (2014), 663–673. https://doi.org/10.1177/0734282914522200
[78] Sangseok Lee and Dong Kyu Lee. 2018. What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology* 71 (10 2018), 353–360. Issue 5. https://doi.org/10.4097/kja.d.18.00242
[79] Shing On Leung and Meng Lin Xu. 2013. Single-Item Measures for Subjective Academic Performance, Self-Esteem, and Socioeconomic Status. *Journal of Social Service Research* 39 (07 2013), 511–520. https://doi.org/10.1080/01488376.2013.794757
[80] Rensis Likert. 1932. A TECHNIQUE FOR THE MEASUREMENT OF ATTITUDES. *Archives of Psychology* 22 140 (1932), 55–55.
[81] Theodore M. Madden and Frederick J. Klopfer. 1978. The "Cannot Decide" Option in Thurstone-Type Attitude Scales. *Educational and Psychological Measurement* 38, 2 (1978), 259–264.
[82] 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
[83] 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
[84] Stephanie M. Merritt, Alicia Ako-Brew, William J. Bryant, Amy Staley, Michael McKenna, Austin Leone, and Lei Shirase. 2019. Automation-induced complacency potential: Development and validation of a new scale. *Frontiers in Psychology* 10, FEB (2019), 1–13. https://doi.org/10.3389/fpsyg.2019.00225
[85] Nicole Mirnig, Astrid Weiss, Gabriel Skantze, Samer Al Moubayed, Joakim Gustafson, Jonas Beskow, Björn Granström, and Manfred Tscheligi. 2013. Face-to-face with a robot: What do we actually talk about? *International Journal of Humanoid Robotics* 10, 1 (2013), 1350011. https://doi.org/10.1142/S0219843613500114
[86] Ranjeev Mittu, Donald Sofge, Alan Wagner, and W. F. Lawless. 2016. *Robust intelligence and trust in autonomous systems.* Springer New York, NY. 1–270 pages. https://doi.org/10.1007/978-1-4899-7668-0
[87] Pam Moule. 2015. *Making Sense of Research in Nursing, Health and Social Care.* SAGE Publications Ltd.

[88] 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

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

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

[91] Takumi Ninomiya, Akihito Fujita, Daisuke Suzuki, and Hiroyuki Umemuro. 2015. Development of the Multi-dimensional Robot Attitude Scale: Constructs of People's Attitudes Towards Domestic Robots. *International Conference on Social Robotics* 1 (2015), 482–491. https://doi.org/10.1007/978-3-319-25554-5

[92] Tatsuya Nomura, Takayuki Kanda, and Suzuki Tomohiro. 2006. Experimental Investigation into Influence of Negative Attitudes toward Robots. *AI & SOCIETY* 20, 2 (2006), 138–150.

[93] Tatsuya Nomura, Keisuke Sugimoto, Dag Sverre Syrdal, and Kerstin Dautenhahn. 2012. Social acceptance of humanoid robots in Japan: A survey for development of the frankenstein syndorome questionnaire. *IEEE-RAS International Conference on Humanoid Robots* (2012), 242–247. https://doi.org/10.1109/HUMANOIDS.2012.6651527

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

[95] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human Computer Studies* 112, July 2017 (2018), 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[96] Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois Press.

[97] Andreea Peca, Mark Coeckelbergh, Ramona Simut, Cristina Costescu, Sebastian Pintea, Daniel David, and Bram Vanderborght. 2016. Robot Enhanced Therapy for Children with Autism Disorders: Measuring Ethical Acceptability. *IEEE Technology and Society Magazine* 35, 2 (2016), 54–66. https://doi.org/10.1109/MTS.2016.2554701

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

[99] Astrid Rosenthal Von Der Pütten and Nikolai Bock. 2018. Development and Validation of the Self-Efficacy in Human-Robot-Interaction Scale (SE-HRI). *ACM Transactions on Human-Robot Interaction* 7, 3 (2018), 30 pages. https://doi.org/10.1145/3139352

[100] 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. *Structural Equation Modeling: A Multidisciplinary Journal* 5511 (2009), 99–117. https://doi.org/10.1207/s15328007sem1301

[101] Er B Ravinder and A B Saraswathi. [n.d.]. Literature Review Of Cronbachalphacoefficient (A) And Mcdonald's Omega Coefficient (Ω). ([n.d.]).

[102] Tenko Raykov and George A Marcoulides. 2011. *Introduction to psychometric theory*. Routledge.

[103] Gary B Reid, Scott S Potter, and Jeine R Bressler. 1989. Subjective Workload Assessment Technique (SWAT): A User's Guide. *ARMSTRONG AEROSPACE MEDICAL RESEARCH LABORATORY* (1989), 115. https://doi.org/45433-6573

[104] John Robinson. 2014. *Faith in People*. Springer Netherlands, Dordrecht, 2151–2152. https://doi.org/10.1007/978-94-007-0753-5_989

[105] John R Rossiter. 2002. The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing* 19 (2002), 305–335.

[106] Julian B Rotter. 1967. A new scale for the measurement of interpersonal trust. , 651–665 pages. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

[107] Miguel A. Salichs, Shuzhi Sam Ge, Emilia Ivanova Barakova, John-John Cabibihan, Alan R. Wagner, Álvaro Castro González, and Hongsheng He (Eds.). 2019. *Social Robotics - 11th International Conference, ICSR 2019, Madrid, Spain, November 26-29, 2019, Proceedings*. Lecture Notes in Computer Science, Vol. 11876. Springer. https://doi.org/10.1007/978-3-030-35888-4

[108] Peter Samuels. 2016. Advice on Exploratory Factor Analysis. *Centre for Academic Success, Birmingham City University* June (2016), 2. https://doi.org/10.13140/RG.2.1.5013.9766

[109] 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

[110] 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 *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 43–52. https://doi.org/10.1145/3371382.3380739

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

[112] 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. *Psychological Assessment* 31, 4 (2019), 557–566.

[113] 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

[114] Bala Subramanian. 2012. Likert Technique of Attitude Scale Construction in Nursing Research. *Asian J. Nursing Edu. and Research* 2 (06 2012), 65–69.

[115] 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

[116] 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

[117] 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

[118] 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.

[119] Andrew J Vickers. 2019. Comparison of an Ordinal and a Continuous Outcome Measure of Muscle Soreness. *Int J Technol Assess Health Care* 4, 1999 (2019), 709–716.

[120] Alan Wagner, David Feil-Seifer, Kerstin Haring, Silvia Rossi, Thomas Williams, Hongsheng He, and Shuzhi Ge. 2020. *Social Robotics 12th International Conference, ICSR 2020, Golden, CO, USA, November 14–18, 2020, Proceedings: 12th International Conference, ICSR 2020, Golden, CO, USA, November 14–18, 2020, Proceedings.* https://doi.org/10.1007/978-3-030-62056-1

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

[122] Bert Weijters, Elke Cabooter, and Niels Schillewaert. 2010. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing* 27, 3 (2010), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

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

[124] Nahathai Wongpakaran and Tinakon Wongpakaran. 2013. Reliability Analysis : Its Application in Clinical Practice. *Chiang Mai University, Thailand* (2013).

[125] 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

[126] Rosemarie E. Yagoda and Douglas J. Gillan. 2012. You Want Me to Trust a ROBOT? The Development of a Human-Robot Interaction Trust Scale. *International Journal of Social Robotics* 4, 3 (2012), 235–248. https://doi.org/10.1007/s12369-012-0144-0

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

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