Choosing the "perfect" scale: a primer to evaluate existing scales in HRI

LAURA SAAD, Naval Research Laboratory
EILEEN ROESLER, George Mason University
ELIZABETH K. PHILLIPS, George Mason University
J. GREGORY TRAFTON, Naval Research Laboratory

Scales are commonly employed in HRI research, yet due to its multidisciplinary nature, many in this community lack direct training in psychometrics. This poses challenges for appropriate scale selection and accurate assessments of reliability and validity. We provide a primer to empower researchers without scale development expertise to assess scale quality efficiently. The guideline provides high-level questions and examples to help the reader make confident evaluations of existing scales in HRI. The guideline is then used to evaluate the Godspeed and Robotic Social Attributes Scale (RoSAS). RoSAS is found to be adequately validated, while Godspeed warrants further investigation before it should be used in HRI contexts. The paper concludes by offering advice on the use of custom scales and provides references for further enhancing expertise in this domain.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**; • **Applied computing** → **Psychology**.

Additional Key Words and Phrases: Scale Development, Psychometrics, Human-Robot Interaction, Questionnaires

ACM Reference Format:
Laura Saad, Eileen Roesler, Elizabeth K. Phillips, and J. Gregory Trafton. 2024. Choosing the "perfect" scale: a primer to evaluate existing scales in HRI. 1, 1 (October 2024), 23 pages. https://doi.org/XXXXXXX.XXXXXXX

1 NAVIGATING THE CHALLENGES OF SCALE DEVELOPMENT AND SELECTION

Measurement is a fundamental aspect of scientific research. Scientific discovery depends upon accurate and reliable measures and the development of such measures requires a principled approach. In human-robot interaction (HRI) research, scales (also known as rating scales or questionnaires), play a crucial role by providing insights into user perspectives and contributing to advancements in robotics and related technologies. Collectively, from 2015 to 2021 the ACM/IEEE International Conference on HRI and the IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) published 1464 papers [82] and of those 61% (889 papers) included scales.

The HRI and RO-MAN community is composed of roboticists, computer scientists, psychologists, cognitive scientists, designers, linguists, and engineers, among others [29]. Many of these disciplines do not provide direct training in psychometric theory. Psychometric theory is the scientific study

Authors' addresses: Laura Saad, laura.s.saad.ctr@us.navy.mil, Naval Research Laboratory; Eileen Roesler, eroesle@gmu.edu, George Mason University; Elizabeth K. Phillips, ephill3@gmu.edu, George Mason University; J. Gregory Trafton, greg.j.trafton.civ@us.navy.mil, Naval Research Laboratory.

ACM XXXX-XXXX/2024/10-ART
https://doi.org/XXXXXXX.XXXXXXX

of testing, measurement, and assessment in the social and behavioral sciences. This lack of training in the HRI community[1] makes it likely that researchers use scales without having the expertise to evaluate whether the scale has been adequately developed and validated. This potential gap, between extensive usage and lack of knowledge, calls for the importance of gaining (at least some) expertise in scale design, development, selection, and evaluation. These skills are particularly relevant to those aiming to advance their research within the HRI community, but also for those who value reproducible research [50]. Developing the ability to critically analyze existing scales will become a necessary tool in the proverbial toolbox if the HRI community wishes to achieve more valid and reproducible results and ensure continued impact in both academic and applied contexts.

Developing a well-designed scale not only takes a substantial amount of time but also requires both art[2] and science [6, 23]. The field of psychometric theory, originating with William Cattell's *Mental Tests and Measurements* [17], contains over a century's worth of research with innumerable articles and textbooks detailing complex methods of analysis aimed at optimizing the process of scale development and validation. Gaining expertise in this field is not a simple task (and will not be accomplished by simply reading one paper) though it is quite easy to implement a scale into a project. This ease of implementation combined with a lack of expertise on scale development and validation makes the HRI community particularly vulnerable to pitfalls associated with measure creation, selection, and use. For example, not understanding how scales are tied to the constructs (sometimes also referred to as "domains"[3]) they measure or to a broader psychological theory can be problematic. This is because the lack of understanding limits the interpretation of the outcomes given by the scale [72]. Additionally, without some knowledge regarding the best practices in psychometric theory, simple mistakes and assumptions can propagate easily leading to work that is potentially uninterpretable. Providing this community with the ability to identify when a scale has not been properly developed will help prevent the proliferation of scales that do not actually measure their stated construct. Avoiding this immense waste of resources should be an appealing goal to the entire HRI community and constitutes a partial motivation for this paper.

The aim of this tutorial is to equip individuals in the HRI community, specifically those without expertise in scale development, with tools to critically assess whether a scale has been adequately tested and validated and to be able to do so relatively quickly. Part two provides a step-by-step guideline on the basics of scale development. The guideline aims to provide the reader with high-level conceptual information that can help them identify minimum criteria in the scale development process and determine the quality of existing scales. Part three applies the guideline to evaluate two frequently cited scales in HRI. Part four provides advice for those interested in using customized scales in their research. Importantly, this guideline does not provide enough information for readers to develop new scales themselves. This information can be found in part five, which concludes the paper with suggestions for future work and some additional resources for those interested in a more in-depth understanding of scale development.

[1]We label these individuals as "non-expert" simply because we assume they are not experts in the field of psychometric theory. The term "non-expert" does not imply that we assume they are "non-experts" in HRI. Though one may identify as a non-expert, they may still incorporate psychometric measures in their research. It is to this group of individuals – non-experts in psychometric theory with an interest in including scales in their research – which we direct the advice and guidelines detailed in this paper.

[2]This claim, that scale making requires "art", stems from the author's (of this paper) previous experience with the scale development process. Specifically, the fact that some steps of the process require creativity, particularly those steps where there are not necessarily steadfast rules or procedures (e.g., item development).

[3]The terms construct, domain, and latent variable are frequently used interchangeably in the psychometric literature and all refer to the same thing—the unobservable behavior, attitude, or attribute that is being measured.

2 GUIDELINES FOR CHOOSING THE "PERFECT" SCALE

The term "scale" refers to any survey that measures a behavior, attitude, or other latent construct that isn't directly observable. Selecting the right scale for a research project is a balancing act. For example, there must be consideration of the research goal, the method of implementation (e.g., online or in-person, between- or within-subjects), and the timeline of the project. There are three possible scenarios that a researcher may find themselves in when considering incorporating a scale into their research project: 1) creating and validating a brand new scale, 2) finding and evaluating the validity of an existing scale, or 3) creating a custom scale[4].

Developing a new scale might be required in cases where the construct of interest has been difficult to measure in the past. It may also be required in cases where the construct has never been measured before. However, as previously mentioned, the process of developing and validating a brand new scale can be challenging for non-experts and the details for conducting this type of research are beyond the scope of this paper (though see part five for some references). Therefore, a reasonable first step is to determine whether or not a scale already exists that measures the attitude, attribute, or behavior of interest.

This paper aims to help the reader select the appropriate scale, evaluate whether it was developed adequately, and determine its validity as a measure. Depending on what the research aims to measure, it is possible there may be multiple scales available. Therefore choosing the "perfect" scale will depend not only on the research goals but also on how psychometrically valid the scale is compared to others. Only in the event that a previously validated scale is not available should the reader proceed to develop a custom scale. Part four of this paper provides information and advice for those who plan to develop a custom scale for their research projects.

The proceeding subsections detail a guideline for choosing the "perfect" (or at least the current best) scale from existing and previously published scales in HRI. The guideline consists of 13 high-level questions that equip the reader with the skills to engage in the critical analysis of a scale before they decide to incorporate it into their study. These questions were synthesized from other sources [6, 23] and the authors' (of this paper) experience analyzing and developing scales.

The authors would like to note that many of the guideline items include recommendations for minimum acceptable criteria. Where possible we provide citations for recommendations with exact values (e.g., Cronbach's alpha). These recommendations can and should be interpreted as heuristics. We acknowledge that the heuristics we provide here are not perfect and we do not encourage the reader to discard a scale simply because it does not meet a specific threshold that has been suggested in this paper.

We also acknowledge that different researchers will use slightly different heuristics across studies and development methods. This is perfectly acceptable in scientific research so long as adequate motivation or rationale is provided. Researchers with more experience in psychometric theory may be able to approach the scale development process with more nuance than a non-expert. For example, consider the use of $p < 0.05$ as a heuristic for "significance" in science more broadly. Though this threshold is far from perfect [32, 66, 74, 76] it can be used as a guard rail to aid researchers in the interpretation of results. An experienced researcher can recognize that the 0.05 threshold is arbitrary and that a p-value of 0.055 is not meaningfully different from (in terms of the interpretation) a p-value of 0.045; both values indicate that the probability of observing the current result is unlikely if the null hypothesis is true. Similarly, in the context of scale development, a ω value of 0.65 is not meaningfully different from an ω value of 0.75; both values suggest that there are reasonably high levels of internal reliability in the measure. When possible, we recommend that the reader considers approaching these heuristics with similar nuance. However, as the target

[4]A custom scale is any scale that has not been validated. See part four for more details.

audience for this paper are researchers without expertise in psychometric theory, we recommend using these criteria as a first step for understanding basic scale development criteria.

This guideline is separated into the three main stages of the scale development process: item development, scale development, and scale evaluation. At each stage, this paper identifies the minimum requirements a "good" scale must meet and poses 2-3 questions. These questions can be used to guide the reader to determine whether the scale has met the minimum requirements to be considered adequately validated. Each subsection includes a brief description of the questions as well as details regarding their relevance and importance to the scale development process.

2.1 Stage 1: Item Development

Item development refers to the process by which the items of the scale are created. Each item is intended to capture the attitude or behavior (i.e., construct) of interest either in part or in full. Items often take the form of direct questions, directives, or statements about their underlying construct. For example, a scale measuring trust might ask participants to respond to the item, "I trusted that the robot was safe to cooperate with" by rating the amount they agree with that statement using a response scale ranging from 1 to 5 [18]. The item development stage lists three main questions the reader should ask of the scale they are evaluating. These questions can help the reader ensure that the item development process was completed adequately and that the construct the scale is intended to measure is appropriately defined.

Question 1: Is the construct that the scale is attempting to measure defined clearly somewhere in the paper? The first step of a typical scientific research endeavor is to clearly state the topic of interest. In psychometric theory, this is referred to as identifying the construct of interest. A construct refers to the unobserved (i.e., latent)[5] behavior, attitude, or attribute that is the target of the study [7, 44]. Unobserved in this context simply refers to a type of construct that exists in the mind of the participant and cannot be directly observed. Measurement scales are typically developed to measure a latent construct that can be inferred from participant's responses to scale items [28, 61]. An example of a latent construct relevant to the HRI might be perceived agency [71]. The perception of agency of another entity cannot be directly measured, as it exists only in the mind of the participant, and therefore must be inferred. Therefore, the perception of agency is considered a latent construct. The spectrum of possible latent constructs that can be measured is expansive and, for example, can range from performance [36], to perceptions [55], to preferences [11].

After identifying the construct the reader must next determine whether or not it has been clearly defined. A clear definition is one that is precise, unambiguous, and completely explains the construct. The definition should not only state what the construct is but also what it isn't [46]. An example of a good definition is from Lee and See [48] who defined trust in automation as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." This definition is good because it is specific and makes clear what aspects the scale developers deemed to be important to the construct of trust (i.e., uncertainty and vulnerability). The definition also provides testable predictions regarding the relationship between trust and other related constructs (e.g., safety) which is an important component in a good definition. Note that [48]'s definition is quite different from [55]'s; this difference is acceptable and good for growth in the field. The reader can select the construct definition and corresponding scale that aligns most closely with their research goals.

There are two approaches that can be used to develop a precise definition of a construct: theory- or data-driven. In a theory-driven approach, a clear definition is synthesized from the existing literature at the start of the scale development process [56]. Ideally, a precise definition that is

[5]The terms "unobserved" and "latent" are often used interchangeably.

agreed upon will already exist in the literature in which case the theory-driven approach is the best avenue forward. For example, Malle and Ullman [55] used a theory-driven approach to determining the construct of trust as "the dyadic relation in which one person accepts vulnerability because they expect that the other person's future action will have certain characteristics; these characteristics include some mix of performance (ability, reliability) and/or morality (honesty, integrity, and benevolence)" (p. 15). However, it is not uncommon to find that a new or more precise definition is needed. If this is the case, the paper should include a brief review of the literature, and previous definitions if necessary, along with the newly proposed definition of the construct that the scale aims to measure. In the case where no agreed-upon theoretical framework exists it is also appropriate to proceed using a data-driven approach.

The data-driven approach can be thought of as a bottom-up process where the researcher is agnostic regarding the latent dimensions that underlie the construct and instead incorporates a wide variety of items that could capture the construct of interest. For example, in their efforts to investigate the dimensions underlying the mind perception in others, Gray et al. [30] included a range of items that captured many different potential underlying dimensions and then evaluated those dimensions post hoc to develop their Experience-Agency framework. Other researchers have also used this approach to examine the latent dimensions of mind perception [54, 77]. The data-driven approach is useful as it allows the data to "speak for itself" [77]. In other words, it may allow for any unexpected structure in the data to be revealed more easily than if a structure was imposed onto it *a priori*.

Either the theory- or data-driven approach can be used, where appropriate, though both should lead to the same destination: a clear definition of the construct. A good definition can help the reader determine whether the scale is appropriate for their purposes, i.e., answering the question: "Is this measuring what I want it to?" For example, a researcher interested in measuring trust in HRI contexts will likely not want to use a measure that has defined trust in other contexts (e.g., Charalambous' scale for use in industrial contexts [18]). It is important that the reader considers the stated definition of the construct carefully before proceeding with the rest of the critical analysis of the scale. It may turn out that the construct, and therefore the scale, is not appropriate for use in their research project. The purpose of evaluating the construct definition is to provide the reader with adequate information to proceed to the next step which is item evaluation.

Take home: A clear and precise definition of the construct is critical. This can be achieved using a theory or data-driven approach depending on whether an agreed-upon theoretical framework of the construct exists or does not, respectively. The reader should ensure that the reported definition of the construct that is being measured by the scale matches the construct of interest for their research project.

Question 2: Is the item generation process discussed (e.g., via a literature review, the Delphi method, crowd-sourcing)? Though not typical in HRI, good practice in scale development is to include some description of the item generation process. Ideally, the paper should start with a large pool of items (usually 2-3x larger than the desired end total) that captures the construct of interest [43, 67]. There are many ways authors of scale development papers may conduct this process. In this tutorial we highlight three common methods: a literature review, the Delphi method, or crowd-sourcing.

A literature review is one of the most common ways an author of a scale development paper might generate items. This process first entails thoroughly reviewing the existing scales (if any exist) as well as the theoretical and empirical literature within the topic of interest. Then the scale developers may identify specific items or phrases that they deem to be directly relevant to the construct they intend their scale to measure. It is probable that scale developers that choose to generate items using this approach will already have some idea or theory about the construct

in advance of the scale development process. This theory will ideally be outlined in the scale development paper to provide clear motivation for both the stated definition as well as the specific items that were included in the initial version of the scale. A good example of a thorough literature review during the process of scale development is the one reported in Malle and Ullman [55].

Another method for item generation is the Delphi method for item development [52]. In this method, experts are recruited to a panel to evaluate whether the scale items adequately capture the construct of interest. The process is iterative and requires each expert be consulted at least twice per item (i.e., once initially and then a second time after considering anonymized feedback from other experts on the panel) [47]. This method is often implemented when the scale developer is not a subject matter expert in the construct they aim intend to measure. Regardless of the expertise of the scale developer, it is generally considered good practice to consult experts in the field during the item development process.

The last method for item generation that we highlight in this tutorial is what we have labeled the "crowd-sourcing" method. Crowd-sourcing refers to a broad variety of methods where lay persons are recruited to consider their interpretation of the construct and provide their explicit thoughts about items or stimuli. Some specific examples of crowd-sourcing include asking participants to sort potential items into categories that are hypothesized factors underlying the construct of interest [55]. This method may also include conducting structured interviews where participants view stimuli (e.g., video of a robot moving through space) are asked questions that pertain to specific components of the construct of interest [18]. Focus groups can also be considered a crowd-sourcing method.

Take home: Look for any information at all about how the items were generated (e.g., via literature review, the Delphi method, or crowd-sourcing).

Question 3: Does the final version of the items capture the construct as it has been defined by the authors? The first step in this process is to determine whether the items are listed verbatim anywhere in the main text of the paper or in the supplementary material. Without this information, the reader cannot determine whether the items are appropriate for their research project.

If the items are listed verbatim it is important to consider whether or not they capture the construct as it has been defined. It is possible the scale has included items that measure a construct that is not part of the stated definition. For example, in the development of their trust scale Yagoda and Gillan [80] reported that their goal was to develop a measure of trust in HRI contexts. However, the majority of the paper was devoted to the determination of dimensions of HRI (e.g., team configuration, context, systems, etc.). These dimensions were not specific to their construct definition of trust and in fact were entirely separate from it. The development of the trust items and dimensions, such as reliability or accessibility, were secondary. Importantly, the authors did not completely capture the construct as it was defined. Therefore, [80] also presents an example where a scale has not captured the construct completely.

At this stage of the evaluation process, it can be helpful for the reader to refer back to the definition of the construct to see if it includes an explanation of the behaviors, attitudes, or attributes that the proposed construct does not encompass. For example, [55]'s stated definition identified performance and morality as the two main components of trust. The precise definition reported in their paper was a result of a thorough literature review combined with a rigorous validation study of their measure of trust. Therefore, it is up to the reader to review the items thoroughly in order to ensure that they encompass performance trust and moral trust and nothing more or less.

This process of checking the items and their match with the reported definition of the construct is related to an important facet within psychometric theory known as construct validity. Construct validity refers to the degree to which a measure actually measures the construct it is proposed to

measure [24]. There are many ways to formally check whether a measure has construct validity (see section 2.3 for more details), however this initial, albeit informal, check by the reader is the first step in the process. It is not uncommon for the initial version of a scale to include items that measure factors that are related to the construct of interest while not actually representing an underlying dimension that contributes to the variability in responses along the construct. In other words, initial versions of scales often include items that don't exactly fit with the construct they aim to measure. Typically during the scale development process, specifically the item-removal stage (see Question nine for more details on this process), items that do not fit within the construct are removed from the final version of the scale. If items are included in the scale that do not appear to be directly related to the construct as it has been defined, that is a sign that the development process of the scale has not been adequately completed. In these cases the reader might also consider searching for another validated scale, if one exists.

Lastly, while ensuring that the items match with the construct definition, the reader should check whether the items in the scale are clear and unambiguous. If the items are not easy to understand that may limit the population the scale is applicable to (e.g., college students). Additionally, ambiguity can increase variability in responses that stem from misunderstanding and not from participant differences across the latent construct. Relatedly, the reader should also ensure the included items are conceptually redundant but not grammatically redundant [26]. This requires evaluating whether the items are simply phrased differently but do not actually capture the full range of the construct of interest. For example, "This robot looks happy" and "The degree to which this robot looks happy" are so grammatically redundant that it is unlikely people would give a different score. This is important to consider as grammatical redundancy increases agreement between items (i.e., reliability) but does not ensure the items capture the entire scope of the latent construct.

Take home: Ensure that the items are related to the reported definition of the construct and also that they are clear and unambiguous.

2.2 Stage 2: Scale Development

De Ayala [25], McCoach et al. [56], Revelle [64] define two approaches to designing and validating a new scale: classical test theory (CTT) and item response theory (IRT). The assumption in CTT is that the participants' responses or overall score on a measure are a linear combination of their true ability plus random error. The goal in CTT is to get as close to the true score as possible by minimizing the noise. IRT, on the other hand, is a more modern method that uses an item-level approach to determining item and person fit within the scale. One type of IRT model that is often incorporated is the Rasch model. The Rasch model prioritizes invariance in measurement [78] and can be thought of as a theory for how the data should be structured which can then be used to identify deviations in observed data. In other words, the Rasch model is a process for fitting data to a model [1, 78].

Though there are many different methods for developing a scale, there are some components of the process that are consistent across methods. We outline these points in the scale development stage of our guideline which includes seven questions. First, the reader should ensure that the sample size of the validation study is appropriate for the scale. This requires consideration of the number of items used in the initial study as well as the type of development method the paper uses [22, 53]. Second, the reader should look for details regarding the analysis of the relationship between the items and the dimensions underlying the construct of interest. The reader should look for reports of some investigation into the number of factors within the construct (i.e., dimensionality) as well as the relationship that exists between those items, factors, and the scale as a whole. The reader should also look for some detailed information about the item removal process. Though this can be done in different ways, depending on the development method used, there should be some

report of the criteria used and how many items were removed before the final scale is reported. Lastly, the reader should determine whether the final version of the scale is reported in the paper.

Question 4: Did the scale developers report the full initial set of items? At this stage, the reader should first determine whether the scale developers have reported the full initial set of items they used when developing the scale. This is important because it will allow the reader not only determine whether the items capture the construct (i.e., guideline question three) but also whether the sample size is large enough to determine a factor structure (i.e., guideline question five). Additionally, the reader can determine if the factor loadings for all the items meet the relevant stated criteria (i.e., guideline question eight) and whether the scale developers removed items appropriately (i.e., guideline question nine). These are all key components of the scale development process and are related to an important tenet of science, replicability. If the initial version of the items is not reported, it will be very difficult for the reader to accurately and confidently determine whether the scale has adequately been developed.

Take home: Ensure that the developers of the scale made the full initial set of items publicly available, either by reporting them in the main text of the paper, in an appendix, or in an online repository.

Question 5: Does the test sample size meet the 10:1 minimum criteria? The sample size for the initial version of the scale should be at least 10:1 [27, 60] participants to items on the scale. However, even larger samples are better [33] particularly if the item reduction procedure is more complex (e.g., parallel analysis [6]). Some authors argue for fewer participants to items (e.g., [37]). This is acceptable in cases where the construct and underlying dimensions are well understood and supported by a developed theory. Strong theory development should ideally lead to strong item to construct relationships. However, HRI is a relatively new field and many of the constructs that are being measured as well as the theories that underlie these constructs are still being developed. Therefore, we recommend the adoption of the 10:1 threshold as a minimum sample size for scale development studies within this research domain.

The recommendation that this criterion be applied to the initial version of the scale is because the item reduction process begins with the data from the initial sample. If this initial sample size is too small (i.e., it does by far not meet the 10:1 criteria) the probability that the observed pattern of results is due to measurement error is increased. It is well-known that larger sample sizes reduce measurement error and it is important for any research study to aim to minimize the amount of measurement error in the data. In scale development, this has many downstream benefits including ensuring the stability of how the items fit to the factor or factors that compose the construct, the replicability of the factor structure, and even potentially in increasing the generalizability of the scale in different contexts (e.g., online vs in-person or in using different stimuli).

We recognize that for practical reasons it is not always possible for a scale development paper to meet this criterion. For example, an initial scale that is 30 items long would require a sample size of at least 300 participants in order to confidently conduct the rest of the scale development process (e.g., factor analysis). This is not always feasible and therefore represents and opportunity for improvement for future studies in HRI. Importantly, however, if a scale does not meet this criterion it does not necessarily mean that it should be discarded. It simply means that the scale in its current form has not met this criterion in the scale development process. It is up to the reader to determine whether the reported sample size is acceptable for their purposes (e.g., there is a difference between collecting a sample of 10 and a sample of 70 when developing a scale with 10 items) and if they are comfortable with the conclusions drawn from the development process as a result.

Take home: Sample sizes for scale development studies should follow the 10:1 (people to initial number of items) rule though more participants is considered a positive feature. This rule pertains to the initial set of items, not the final version of the scale.

Question 6: Did the scale developers perform an EFA, PCA, Rasch analysis, or similar test to determine the item to factor relationship? Determining the different factors or characteristics that compose a construct as well as how those factors relate to each other is an important tenet within scientific research. In psychometric theory, this requires investigating how the items capture the underlying structure of the construct of interest. The assumption is that the observed data pattern is a result of some relationship between the factors that are not directly observable. Factors we cannot directly measure or observe are referred to as latent factors and some examples of latent factors of interest to the HRI community include trust, confidence, or perceived agency.

There are many ways to investigate the relationship between items, factors, and the construct of interest. This can be completed using methods such as exploratory factor analysis (EFA), principal components analysis (PCA)[6], or the Rasch model. There is a huge corpus of scholarly works devoted to scale development using these methods. While it is beyond the scope of this article to describe all the details about evaluating these methods, we highlight these analysis methods to act as heuristics that the non-expert reader can search for when determining whether a scale has been developed adequately. (The inclined reader can learn more about factor analysis and other methods of investigating the relationship between items and dimensions in the following resources: [28, 43, 61, 63, 64, 78].) What the reader should ensure is that at the very least the study should include some description of how the scale developers determined the number of dimensions. We refer to this step in the process as the scale development method and can include mention of conducting an EFA, PCA, or Rasch analysis.

Take home: There are many methods that can be used to determine the underlying factor structure of the construct of interest. The reader should determine whether the scale developers report using at least one scale development method (such as EFA, PCA, or Rasch) in their paper.

Question 7: Did the scale developers describe how they determined the number of factors? Determining the number of factors within a construct is not always a straightforward process. A construct can be unidimensional (i.e., consisting of only one factor) or multi-dimensional (i.e., consisting of more than one factor). A unidimensional scale measures the construct along a single range from low to high. For example, the perception of agency scale [71] is a unidimensional scale. More common are multi-dimensional scales like trust [18, 55, 73], negative attitudes towards robots [59], or perceived morality [2]. Multi-dimensional scales measure a construct along different dimensions and are then typically combined for an overall measure of the entire construct. The goal of the factor extraction process is to determine the minimum number of factors that are necessary to describe and interpret the data.

There are number of different ways to determine the number of factors and each way is dependent on the scale development method used. Some potential methods that scale developers might report using are scree plots, parallel analysis of random data, statistical tests such as chi-squared test of residuals, very simple structure (VSS), or minimum average partial (MAP)[7]. The technical details of each of these tests is beyond the scope of the paper. The reader only needs to determine whether a method to determine factors was reported by those developing the scale.

[6]Note that the authors recommend PCA should not be used for scale creation (NEED REF).

[7]These methods are specific to those scale development papers using EFA or PCA as the Rasch analysis method should only be used on unidimensional data [8, 31].

Take home: Constructs can be unidimensional or multidimensional. The reader needs to determine if the scale developers reported exactly how they used the method they described (EFA, CFA, Rasch) to determine the number of factors (i.e., dimensions) exist within the construct.

Question 8: Did the scale developers provide factor loadings (EFA/CFA) or item fits (Rasch) of all items? After determining the number of dimensions, the paper should report the relationship between the items and the construct (including the dimensions). For exploratory factor analysis, this includes reporting factor loadings for each item. Factor loadings represent how well each item correlates with all the other items in that dimension, or how much variance or covariance each latent factor is capable of explaining. Higher values are better with a minimum of 0.6 [60]. Once the factor loadings are obtained, they are rotated so that the simplest underlying structure can be revealed. Rotation can either be orthogonal rotation (assuming factors are uncorrelated) or oblique (assuming factors are correlated). Confirmatory factor analysis also provides factor loadings and so the minimum value of 0.6 can also be applied when using this method. Rasch analysis uses outfit and infit measures [8, 79]. Generally, Rasch items show poor fitting items when an outfit is higher than 1.5 [51]. The reader should ensure that the scale developers have explicitly reported using these values in determining the number of factors and the item to factor relationship.

Take home: The reader should look for quantitative values that indicate how the items in the scale relate to the construct of interest. These values can be in the form of factor loadings (if the scale development process used an EFA or CFA) or in the form of infit/outfit values (if using Rasch analysis).

Question 9: Is there a description of the item removal process (e.g., using infit/outfit, factor loading minimum values, or cross-loading values)? Removing items that are not relevant to the domain of interest, or item reduction, is a critical step in the scale development process. It is very likely that the initial set of items in its entirety will either not be appropriate for the construct or they will not be able to capture the full scope of the construct. Having a principled way of removing items that do not fit with the construct is necessary as is the detailed reporting of that procedure.

There are many different ways to remove items and each method depends upon the scale development method. For example, if the scale was developed using factor analysis then the items might be removed based on factor loading values < 0.3 [6] or high cross loading values (e.g., values > 0.4 one two or more factors) of one item across multiple factors [37]. Additionally, those using the Rasch model may remove items according to fit statistic values used to determine item fit to construct, such as infit > 0.6 [31] or outfit < 1.3 [9]. The exact criteria used may vary across publications.

Items can be removed not only due to lack of fit within the domain but also due to redundancy with other items [26]. As mentioned in question three, when an item is redundant that means that there is more than one item that captures the factor to a similar level. This goal is distinct from removing items due to lack of fit with the domain since, in the former case, the item is measuring a different construct which can add noise to participant responses and mask the true structure of the construct. In the latter case, there is also an increased risk of noise but from an entirely different source. If more than one item is included that captures a similar aspect of the construct, it is not necessary to include it in the final version of the scale. Additionally, since shorter scales are often more easily incorporated into research projects, considering and removing redundant items is an additional step that should be reported in the paper. The methods for removing redundant items are again dependent on the scale creation method but well-built scales should report how redundant items were removed.

Regardless of the specific details, the reader should determine whether or not the scale had consistent criteria for this process. Importantly, they should be able to use the information reported in the paper to replicate the process from start to finish in order to be confident that the scale development process was adequately completed.

Take home: Item reduction can be done in a number of ways depending on the scale development method. Additionally, items can be removed for different reasons: lack of fit with the construct or redundancy with other items. What is important for the reader to identify is whether a criteria was reported and if so, whether it was used consistently.

Question 10: Did the scale developers report the complete list of items included in the final version of the scale? Providing the final version of the scale in the publication is critical. This ensures accessibility, replicability, and, importantly, that the scale is used as intended. If the items are not listed verbatim, the chances of future studies incorporating a scale that includes items that do not adequately measure to the construct increases. This is problematic as the inappropriate use of a scale can waste valuable resources, and even potentially lead to false conclusions drawn from faulty data.

The reader should look for a table that lists the items included in the final version of the scale in the main text, in an appendix, or, in rare cases, as a download-friendly document that includes instructions for administration and scoring.

Take home: It is critical that the final version of the scale in the publication is clearly reported to ensure that the scale is used as intended. The reader should look for this information either in the main text of the publication, in an appendix, or in an online repository.

2.3 Stage 3: Scale Evaluation

Scale evaluation occurs after the original scale is created and attempts to answer the following three questions.

Question 11: Did the scale developers include a factor structure test (e.g., second EFA, CFA, DIF, test of unidimensionality if using Rasch, or similar)? After a scale has been created, it is best to determine if the scale has the same factor structure on a different sample. If factor analysis was used to create the scale, it is common to use a confirmatory factor analysis (CFA) to test the factor structure. When using CFA, the latent structure uncovered during the exploratory factor analysis is used as a hypothesized model on a new set of data [81]. To conduct a confirmatory factor analysis, the researcher uses the results of the initial factor analysis as a set of model parameters for a CFA. It is possible to then examine how well the CFA fits the data; most researchers will report a series of fit statistics including Root Mean Square Error of Approximation (RMSEA), Tucker Lewis Index (TLI), Comparative Fit Index (CFI), and standardized Root Mean Square Residual (SRMR) [12, 13, 35], though others can also be used. Each fit statistic has a heuristic value that the CFA should be under (or over). For example, [40] recommended the following thresholds for fit indices: CFI $\geq$ 0.95, TLI $\geq$ 0.95, RMSEA $\leq$ 0.06, and SRMR $\leq$ 0.08. A CFA should thus report some measure(s) of fit and the acceptable range.

Methods of scale creation besides factor analysis typically use alternative methods to determine whether a scale has the same structure. For example, researchers using the Rasch method will typically focus on measurement invariance using Differential Item Functioning (DIF) [9, 78]. DIF examines two different groups of scale-responders (e.g., male/female or old/young or US/Japan) to determine if the model fits the data for both groups equally as well.

Sometimes a CFA or DIF will discover a weakness in the original scale—an item that does not work as well as expected, suggesting that the item should be removed, replaced, or corrected in

some way. In this case, the reader should look for a report of another CFA conducted on a different group of participants which examined the factor structure of the updated scale.

Take home: Check to see if there is a test for factor structure. A confirmatory factor analysis on a new sample or a Differential Item Function (Rasch) are common approaches.

Question 12: Was a measure of reliability (e.g., Cronbach's alpha, McDonald's ω_t or ω_h, Tarkkhonen's Rho) reported? Reliability refers to the principle that a measurement produces similar results under similar conditions[8] and is related to one of the core components of science: replicability. For a scale to be considered adequately developed and validated it must both measure the construct it is intended to measure (i.e., have construct validity) and also do so reliably. In addition, reliability is a starting place for establishing scale validity, as a measure cannot be more valid than it is reliable. This is because validity and reliability go hand in hand. If the scale is a valid measure of the construct but is not consistent, then its validity as a measure will decrease. However, if it is both a valid and reliable measure of the construct then its validity will be improved.

In the context of scale development, an important component of reliability is the internal consistency of the scale. In order to establish internal consistency, the sources of error in a scale must be determined. Omega total (ω_t) and Omega hierarchical (ω_h) are good measures of internal reliability [23, 65]. Reliability measures should be as high as possible. ω_t is a measure of the amount of variance attributable to a general factor (the primary latent variable) and specific factors (items) while ω_h is a measure of the amount of variance attributable to only the general factor. ω_t can be used for both unidimensional and multi-dimensional scales, while ω_h should only be used for multi-dimensional scales [19].

Cronbach's coefficient alpha (α) is another metric that can be used in conjunction with ω_t or ω_h. α represents a measure of how often the items in a scale actually agree on what they are measuring. Therefore, a high α value means that the relationships between the items account for most of the overall variability. α has been critiqued previously [20, 23, 34, 69, 83] for example, because of its dependence on the total number of items or by including items that have grammatically similar wording.

Many researchers use $\alpha \geq 0.70$ as a traditional heuristic. This often-cited standard threshold for reliability comes from Nunnally [60]. Interestingly, a closer inspection of the original text shows that this is in fact a misrepresentation (as has been previously noted in [23]). Nunnally [60] writes:

> In the early stages of research on predictor tests or hypothesized measures of a construct, one saves time and energy by working with instruments that have only modest reliability, for which purpose reliability of 0.70 or higher will suffice... In contrast to the standards in basic research, in many applied settings a reliability of 0.80 is not nearly high enough. (p. 245)

This implies that $\alpha = 0.70$ is a useful starting point but certainly not adequate for applied (or even in some cases basic) research settings. This is particularly relevant to cases where the results will inform decisions that impact society, as is the case with some HRI research. For example, in cases where robots are a critical component during military operations or when introducing a robot to an industrial setting where humans are present. As a result, we suggest a minimum value of ≥ 0.80 for low-stakes research and ≥ 0.90 for high-stakes measures. Therefore, considering all of these points, we recommend the use of ω as a reliability metric in place of, or at least in addition to, α. For both α and ω, higher values are preferred.

[8]During the scale development phase, data collection methods may vary across studies. For example, some studies are conducted entirely online (e.g., [71]) while others are in-person (e.g., [45]). To our knowledge, there is no theory about the relationship between data collection methods and validity, particularly in HRI contexts. The current assumption is that these are all equivalent methods and that the results should be reliable across these different contexts and environments.

Importantly, it is still crucial to report the reliability of a scale in cases where the value is low or below the acceptable threshold. Low reliability may be due to various factors out of the scale developer's control (e.g., the scale may not be the best measure of the construct; participants may not be answering correctly or honestly; stimuli may be out of bounds for the scale). Reporting reliability measures (even when they do not meet the minimum criteria) allows for the potential that future experiments and validation studies can account for these problems.

Take home: There should be some test of the scale's reliability in the paper. This can be completed using metrics such as ω_t or ω_h in addition to Cronbach's coefficient α. A reasonable minimum threshold for reliability, particularly when using α, is a value ≥ 0.80. There is no official cut-off value when using ω_t or ω_h, though higher values are preferred.

Question 13: Was a test of validity (e.g., predictive, concurrent, convergent, discriminant) reported? Reliability is not the same as validity and both are vital to the scale development process. Validity measures the extent to which the scale actually measures the latent dimension it was developed to evaluate and is a fundamental concept within psychological measurement [24, 57, 63]. The concept of validity can be split into sub-components such as criterion and construct validity [6, 24]. Criterion validity refers to the degree to which there is a relationship between the construct on the current scale and construct on another similar measure or in another context that is of interest to the researcher. Criterion validity further breaks down into predictive and concurrent validity. Predictive validity measures the degree to which performance on the current scale predicts performance on another scale taken at a later time. Concurrent validity measures the degree to which the performance on the current scale relates to performance on a criterion (gold standard) measurement [24]. Typically, the two measures are administered at the same time or consecutively (hence "concurrent"). It is common, however, that no gold standard measure exists making evaluation of concurrent validity impossible [6].

Construct validity on the other hand typically refers to the extent to which the scale measures what it was developed to measure and how much it is associated with other factors within the domain [6, 10]. Construct validity can be measured in many ways [15, 21] though we highlight two common approaches here: convergent validity and discriminant validity. Convergent validity refers to how well the new scale correlates with other variables that are designed to measure similar constructs. Discriminant validity refers to the extent to which the scale is novel. Discriminant validity is measured by analyzing correlations between the measure of interest and other measures that do not measure the same domain or concept [6] where weaker correlations are expected.

The comparison of the scale to others in the field has the potential to offer useful information. Though this comparison is just one avenue to confirm the validity of the scale, it is fairly straightforward to conduct if there are other measures that are related to the one that is being developed or validated.

Lastly, it is often the case that due to practical constraints (e.g., limited time and resources) that a rigorous and formal validation study is not conducted or reported. While we believe formal validation to be a critical step in the scale development process, its absence in the process does not preclude a scale from being used or even from being considered a useful measure. We recommend readers interested in using a scale that has not been formally validated consider conducting a formal validation study (and publishing the results) or explicitly acknowledge this limitation in publications or presentations.

Take home: Look for comparisons of the scale of interest to others in the field and see if there are any relationships that exist. If there is a strong relationship between scales measuring distinct constructs or factors, then more work needs to be done before the scale can be used. If no report of validity has been conducted, then the reader should

explicitly report this limitation when publishing or presenting results. If the reader has the resources, we encourage them to conduct a validation study and publish the results!

2.4 Interim Conclusion

At this point, we hope the reader has developed a basic understanding of the different types of analyses that are part of the scale development and validation process. Here we will briefly summarize the information that was provided in this guideline.

The first stage, item development, involves determining whether the scale measures a well-defined construct. This can be done either by starting with a clear theoretical framework from existing literature or by using a bottom-up approach, where the construct's definition emerges from analyzing the data structure revealed during pilot testing. The reader should also ensure that the item generation process is discussed. Additionally, the reader should pay close attention to the match between the definition and items so as to ensure that the entire construct is captured and no items are incorrectly included in the final version of the scale. This should help the reader be sure that the scale is measuring their desired construct. Item development is a critical step in the process of choosing the "perfect" scale.

The second stage, scale development, delves into the more technical aspects of the process. The first step in this stage is to determine whether the scale developers reported the full initial set of items used in the development process. Then the reader must ensure that the pilot sample was large enough to conduct the appropriate analyses using those items. The ideal sample size is at least 10 participants for every item (e.g., 120 participants for a 12-item scale). The reader should next determine which method was used to determine the underlying factor structure of the construct. Though there are many different and accepted methods for this process, the reader should look for at least one method. The method should detail some explanation of how the number of factors was determined and how the items are related to the factors (sometimes also called dimensions) that make up the general construct of interest. This stage also involves item reduction or removal and the reader should look for details on this process, particularly regarding the threshold or inclusion/exclusion criteria that was used. Lastly, the reader should determine whether the paper has clearly included a list of the items used in the final version of the scale.

The third and final stage, scale evaluation, consists of reliability and validity checks. First, the reader should look for a test regarding the consistency of the reported factor structure. Second, the scale should have some acceptable measure of reliability. Ideally, the scale developers will have computed ω_t or ω_h for the scale, depending on the dimensionality, in addition to the typical Cronbach's coefficient α. Lastly, the reader should look for a test of validity.

Armed with this information, the reader should now feel more confident in critically analyzing existing scales and their corresponding validation reports. To further increase this confidence, we next briefly turn to two examples from the HRI literature and apply the guideline to evaluate whether these scales meet the minimum acceptable criteria as has been suggested here.

3 EVALUATING EXISTING HRI SCALES

This section will use the guideline presented in this paper to evaluate two scales that are frequently used in HRI – the Godspeed Questionnaire [5] and the Robotic Social Attribute Scale [16]. We first briefly describe each paper and evaluate each scale in turn according to the guidelines (see Table 1 for a brief summary). The Godspeed questionnaire was developed as a tool to measure commonly used concepts related to the perception of robots in HRI contexts. It consists of five different questionnaires that are assumed to capture the concepts of anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. The Robotic Social Attribute Scale (RoSAS)

was developed to measure the social perception of robots [16]. It consists of three underlying scale dimensions: warmth, competence, and discomfort.

When the guidelines are applied to the Godspeed scale it is clear that it does not meet the acceptable standards to be considered a reliable or valid scale. The Godspeed scale is composed of five different scales that are assumed to capture different dimensions within the broader construct of the perception of robots. Four of these five scales are custom scales that were developed in previous publications to measure distinct constructs [49, 58, 62, 75] and then these custom scales were combined into the larger Godspeed scale. There are several psychometric concerns with this approach. First, many of the details about item construction/removal, factor/dimension identification, and investigations of the relationship between items and factors are relegated to the original publications, making critical analysis of the scale more difficult for the reader as they must track down and apply the guideline to the original validation studies to determine if the scale is useful for their research. Additionally, items that were included in the final version of Godspeed were modified versions of the original items and only α was reported as a reliability measure on the new items. It is not appropriate to assume that large changes to items or scales will have the same psychometric properties as the original scale. This approach also is unfair to the original scale creators: they do not get citations or other types of credit for doing the original work.

To see some of these issues more clearly, we can take the anthropomorphism scale as an example and evaluate the Godspeed version and the original version using the guidelines (see results in Table 2). The Godspeed version of this scale (reported in [5]) meets three of the criteria: adequate construct definition, final version of scale reported, and reliability test results reported. However, the studies that the reliability values came from were detailed in other papers [3, 4] which were not scale development papers. Additionally, only reliability (α) was reported as a metric of scale quality. Since the anthropomorphism items within Godspeed are modified versions of the originals from the "humanlikeness" scale reported in Powers and Kiesler [62] it can be argued that the guideline should be applied directly to that scale. In doing so, we see that some additional criteria were met, specifically that the scale development method was mentioned. However, there are still many critically important details missing from the original paper including a precise definition of the construct as well as a detailed explanation for how the number of factors was determined and how the items related to those factors. At present there is evidence that neither the Powers and Kiesler [62] nor the Godspeed anthropomorphism scale [5] were adequately developed and validated.

Although the customization approaches reported in Bartneck et al. [5] to develop the Godspeed scale are not ideal, does not mean that it is inappropriate to use the scale in all cases. A crucial component missing from the original publication was that the Godspeed scale was not assessed as a whole. There was no report of how well all of the items fit together across the scales to measure the perception of robots. There were no reported reliability or validity tests of the scale as a whole in any context within the original publication [5] nor were there any citations to studies where the scale was separately validated or assessed. This lack of information makes it impossible for the reader to evaluate whether the scale is a useful or even adequate measure for their research purposes. If the authors of the Godspeed scale had included these analyses initially it may have been easier for the readers to assess its adequacy as a measure of perception of robots. Additionally, it may have been clear at that point that more scale development was needed as it has since been shown that some of the scales included within Godspeed are not adequate measures of the proposed constructs (e.g., see [16, 39, 42]). Thus, based on the results from this evaluation, the Godspeed questionnaire should not be considered a valid measure of user perception of robots in HRI settings until further validation studies are conducted (though see [39] for a validated scale measuring some of the constructs of interest in Godspeed).

Table 1. Applying the Guideline to Two HRI Scales

Stage	Question	Godspeed	RoSAS
Item Development	1. Construct defined?	✓	✓
Item Development	2. Item generation process discussed?	×	✓
Item Development	3. Final items capture the construct?	×	×
Scale Development	4. Full initial set of items reported?	×	✓
Scale Development	5. Person:initial items 10:1?	×	×
Scale Development	6. EFA, PCA, Rasch to determine item:factor?	×	✓
Scale Development	7. Factor extraction method discussed?	×	✓
Scale Development	8. Factor loadings or item fits provided?	×	✓
Scale Development	9. Item removal process described?	×	✓
Scale Development	10. Final version of scale reported?	✓	✓
Scale Evaluation	11. Test for factor structure?	×	✓
Scale Evaluation	12. Reliability reported?	✓	✓
Scale Evaluation	13. Validity reported?	×	✓

When the guideline is applied to the RoSAS we see that it meets almost all of the guideline criteria. The paper reports a clear definition of the construct of social perception of robots as consisting of three factors, two of which were determined via a literature review, and the third, discomfort, was determined as a result of the scale development process. The item generation process was described in detail throughout three studies and the items evolved from the original items used in the Godspeed questionnaire to items that were found to more accurately reflect the underlying factors within the construct. Additionally, in study 2 the authors reported including many additional items (83 total) to ensure that the full range of the construct was captured. However, they do not report those items making it difficult to determine whether they adequately capture the construct. Notably, the construct of "social attributes of robots" is quite large and therefore it is unlikely that the items fully capture the construct as it is defined in the paper. Additionally, the sample size for the pilot studies was not adequate per the 10:1 guideline criterion. Item removal and analysis of factors/dimensions to items were described in studies 2 and 3 for the factors warmth and competence (study 2) and discomfort (study 3). Factor loadings from exploratory factor analyses were the primary way by which both of these analyses were conducted. Results from exploratory factor analysis, reliability analysis, as well as the validation study (study 4) were included in the paper as well which allows for the RoSAS to meet the guideline criteria for reliability and validation. Therefore based on this analysis, the reader can consider RoSAS a valid scale and feel confident incorporating it into their research.

Our comparison of two extensively utilized scales within the HRI community demonstrates that the frequency of usage doesn't necessarily correlate with quality. We aim for these guidelines to empower researchers to select the most suitable scale for their research question, rather than defaulting to the most commonly employed one. In certain instances, the search for existing scales may necessitate the adaptation of scales, paving the way for the subsequent section.

4 ADVICE FOR USING CUSTOM SCALES

What should a researcher do when they need to measure a latent construct? The best and strongest idea is to find and use a scale that has already been created and use the guideline to determine whether it has been psychometrically validated as described in this article. If this process was

Table 2. Applying the Guideline to Two Anthropomorphism Scales

Stage	Question	Bartneck et al.	Powers and Kiesler
Item Development	1. Construct defined?	✓	×
Item Development	2. Item generation process discussed?	×	×
Item Development	3. Final items capture the construct?	×	×
Scale Development	4. Full initial set of items reported?	×	✓
Scale Development	5. Person:initial items 10:1?	×	×
Scale Development	6. EFA, PCA, Rasch to determine item:factor?	×	✓
Scale Development	7. Factor extraction method discussed?	×	×
Scale Development	8. Factor loadings or item fits provided?	×	×
Scale Development	9. Item removal process described?	×	×
Scale Development	10. Final version of scale reported?	×	✓
Scale Evaluation	11. Test for factor structure?	×	×
Scale Evaluation	12. Reliability reported?	✓	✓
Scale Evaluation	13. Validity reported?	×	×

done correctly, the scale should have a strong majority of checks using our approach. However, sometimes a needed scale may be too niche and, due to practical constraints, the researcher might not have the time or expertise to create and validate a new scale. The worst thing a researcher could do at this point is to haphazardly combine individual or all concepts of interest without a systematic approach, resulting in the creation of either a single item or a potpourri of items assumed to measure relevant aspects of a construct. The most common and accepted approach is to generate a *custom scale*. A custom scale is any scale that has not been validated.

There are many ways that a researcher could go about creating a custom scale. A researcher may take a subset of items from an existing complete scale or subscale[9]. This frequently occurs because the original scale is too long and the researcher assumes that the shorter scale will be just as good as the full scale. Unfortunately, removing items increases the possibility that the full spectrum of the domain of interest is no longer represented. This is problematic as it can affect the validity of the measure and researchers can not claim the smaller scale has all the features of the validated scale[10].

Another way that researchers may create a custom scale is to make up their own items based on the literature, their own understanding, and perhaps even from other existing scales. Researchers may also greatly change the wording of an existing scale. Small changes are usually considered acceptable – tense, gender, changing the word "automation" to "robot" [41], etc. are all acceptable, whereas large changes to wording or phrasing are not.

[9]Subscales refer to complete sets of items that load onto one factor in an existing validated scale. For example, the competence subscale in the RoSAS consists of six items that are related to the intelligence or ability of the robot [16]. Including only a subscale in a study is completely acceptable.

[10]The authors of this paper would like to note that it is sometimes possible to remove items from a scale without significant negative impact to the ability of the scale to measure the construct. In an ideal world, if a scale contains an item that is perfectly related to the construct then it is acceptable to use that single item to measure the construct. However, the use of a single item to measure a construct is still controversial in the literature [23]. Removing items is a nuanced process that requires expertise and knowledge that a non-expert in psychometric theory may lack. This motivates us to caution against the removal of items unless further testing and validation is completed. Those with the proper training should feel free to customize a scale and report the changes as appropriate.

Changing the response scale (e.g., range of a Likert scale) of an existing scale is not recommended[11]. In some cases, it can change the meaning of the scale (e.g., converting a scale with an odd number of response categories with a midpoint to an even number of response categories would force a choice in response by the participant.). In some analysis methods (e.g., Rasch), the exact response range is critical to generating acceptable data. These types of adjustments, while seemingly arbitrary, can change the fundamental structure of the scale. Therefore we do not recommend making adjustments to response scales unless the intention is to conduct further scale development.

Another potential situation that the researcher may find themselves in is that they have found a scale that has been developed and validated adequately but it is not a valid measure in their native language. For example, a validation on a translated version of the scale may not have been conducted or specific items in the original version of the scale may have no direct translation into the reader's native language. In cases like this, there are a few paths forward. In an ideal case, the researcher should translate the scale to their native language, collect data on the translated version, and conduct a confirmatory factor analysis using the factor structure from the originally published paper[12]. The reader should then report the translated scale and the CFA fit indices (e.g., RMSEA, TLI, CFI) in a publication, even if the the model fits do not meet the minimum criteria.

However, practical constraints may prevent a researcher from conducting this type of analysis. In that case, the researcher might simply translate the items to their native language and use it. When embarking on this path, it is important that the researcher explicitly reports which language the scale has been translated from as well as a report of reliability (e.g., ω or α), at minimum. Additionally, the authors must also report a verbatim list of the translated items (in their native language). Including the translation used is critical to ensure the replicability of the translated version of scale and increase comparability across studies. Note that there are concerns with this method as translating scales can be a difficult and error-prone process. For example, different cultures may have different understandings or norms, and idiomatic wording in either language can change the meaning of items and the scale. Therefore, we recommend researchers proceed with caution and be as transparent as possible when translating and reporting data from translated scales.

If a researcher does decide to generate a custom scale, we recommend that the researcher be explicit that the scale is a custom scale. We suggest generating 4-6 items that the researcher believes best capture the latent variable of interest then describe the modifications or item creation method process and report reliability measures. The danger of not performing these minimal steps is that other researchers may assume that because your research got published, your scale must be valid; this is certainly something we want to avoid. If future researchers want to use your scale, that is completely acceptable, but they would also need to be explicit that the scale is a custom scale and has not been psychometrically validated.

Lastly, sometimes given the nature of the study, it is not possible to include a lengthy scale that consists of a series of well-validated items. This could be due to a time constraint within the experiment, repeated measurement designs, or budget and funding restrictions. In cases like this, we recommend including as many items as possible that can adequately measure the construct. If the situation allows for the inclusion of only a single item without conducting the adequate prior validation then we recommend the researcher clearly acknowledge the limitation of the measure in the paper. They should also suggest (or conduct) a follow-up study which includes a longer, validated scale to which the single item can be compared.

[11]Here again, the authors of this paper would like to note that while it is possible to change the endpoints of the scale it makes interpretation of results in the context more complicated and makes the comparison between different studies challenging. Therefore for the purposes of the target audience we caution against the modification of response scales.

[12]For the inclined reader, see [14, 38, 70] for more details on this process and some examples in HRI contexts.

5 CONCLUSIONS

Our aim for this primer was to provide a straightforward reference for those who need to think critically about scales and the scale development process. Minimally, we hope that after reading through this paper those in the HRI community, specifically those without direct training in psychometric theory, are better equipped to critically analyze and implement existing scales into their research. On the surface, implementing a scale is deceptively simple and easy. However, determining whether that scale has been designed and developed appropriately is not a simple process. We hope that our guideline has provided the information necessary to determine whether that scale was developed and validated appropriately.

If the scale was constructed well and measures your domain of interest, it should not need to be adapted much (or preferably at all) to fit the needs of the study. If the scale was not adequately validated (i.e., the scale developers only reported $\alpha > 0.70$) there are additional steps that should be taken before incorporating the scale and interpreting its result. This fact remains true whether the scale has been used 3 times or 1000 times. Alternatively, creating a custom scale might be the appropriate course of action in many cases (e.g., due to time constraints, lack of expertise, or lack of existing scales for a robot-related construct). Those researchers creating custom scales should be extremely clear about the modifications made and the motivations for doing so and should do what they can to ensure that the scale adequately measures the construct of interest. This is to ensure reproducibility but also to avoid custom scales that lead to custom scales (and so on endlessly) that are never checked or validated.

If you are considering incorporating a scale into your project it is important to start by tracking down the original validation study of the scale you are interested in using. Then you can use this guideline to determine whether it is a good measure of your domain of interest. If there is no scale that exists, it is likely that you will need to start from the ground up and develop an entirely new measure. Importantly, this guideline should not serve as a reference for developing or validating your own scale. Those interested in learning more about the scale development process can begin with these resources [6, 9, 23, 46, 51, 64, 68, 78, 79].

HRI and robotics research have the potential for a broad impact on society. Maintaining rigor and high-quality standards of our experiments and measures is essential to ensuring the impact we have is a positive one. We hope this paper can contribute to that ideal and serve as a useful reference for any researcher interested in incorporating surveys and scales into their projects, regardless of experience level or field of research.

ACKNOWLEDGMENTS

This work was supported in part by ONR to GT. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the US Navy.

REFERENCES

[1] Vahid Aryadoust, Li Ying Ng, and Hiroki Sayama. 2021. A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. Language Testing 38, 1 (2021), 6–40.

[2] Jaime Banks. 2019. A perceived moral agency scale: development and validation of a metric for humans and social machines. Computers in Human Behavior 90 (2019), 363–371.

[3] Christoph Bartneck, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2007. Is the uncanny valley an uncanny cliff?. In RO-MAN 2007-The 16th IEEE international symposium on robot and human interactive communication. IEEE, 368–373.

[4] Christoph Bartneck, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. My robotic doppelgänger-A critical look at the uncanny valley. In RO-MAN 2009-The 18th IEEE international symposium on robot and human interactive communication. IEEE, 269–276.

[5] 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 (2009), 71–81.

[6] 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 (2018), 149.

[7] Kenneth A Bollen and Rick H Hoyle. 2012. Latent variables in structural equation modeling. Handbook of structural equation modeling 1 (2012), 56–67.

[8] T Bond and C Fox. 2001. Applying the Rasch model. Mahwah, NJ: L.

[9] William J Boone. 2016. Rasch analysis for instrument development: Why, when, and how? CBE—Life Sciences Education 15, 4 (2016), rm4.

[10] Denny Borsboom, Gideon J Mellenbergh, and Jaap Van Heerden. 2004. The concept of validity. Psychological review 111, 4 (2004), 1061.

[11] John Brooke. 1996. Sus: a "quick and dirty'usability. Usability evaluation in industry 189, 3 (1996), 189–194.

[12] Timothy A Brown. 2015. Confirmatory factor analysis for applied research. Guilford publications.

[13] Timothy A Brown and Michael T Moore. 2012. Confirmatory factor analysis. Handbook of structural equation modeling 361 (2012), 379.

[14] Jie Cai, Yuxuan Sun, Chunling Niu, Wei Qi, and Xurong Fu. 2024. Validity and reliability of the Chinese version of robot anxiety scale in Chinese adults. International Journal of Human–Computer Interaction 40, 13 (2024), 3355–3364.

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

[16] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction. 254–262.

[17] James McKeen Cattell. 1948. Mental tests and measurements, 1890. (1948).

[18] George Charalambous, Sarah Fletcher, and Philip Webb. 2016. The development of a scale to evaluate trust in industrial human-robot collaboration. International Journal of Social Robotics 8 (2016), 193–209.

[19] Eunseong Cho. 2022. Reliability and omega hierarchical in multidimensional data: A comparison of various estimators. Psychological Methods (2022).

[20] Eunseong Cho and Seonghoon Kim. 2015. Cronbach's coefficient alpha: Well known but poorly understood. Organizational research methods 18, 2 (2015), 207–230.

[21] Gilbert A Churchill Jr. 1979. A paradigm for developing better measures of marketing constructs. Journal of marketing research 16, 1 (1979), 64–73.

[22] Lee Anna Clark and David Watson. 2016. Constructing validity: Basic issues in objective scale development. (2016).

[23] Jose M Cortina, Zitong Sheng, Sheila K Keener, Kathleen R Keeler, Leah K Grubb, Neal Schmitt, Scott Tonidandel, Karoline M Summerville, Eric D Heggestad, and George C Banks. 2020. From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the Journal of Applied Psychology. Journal of Applied Psychology 105, 12 (2020), 1351.

[24] Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. Psychological bulletin 52, 4 (1955), 281.

[25] Rafael Jaime De Ayala. 2013. The theory and practice of item response theory. Guilford Publications.

[26] Robert F DeVellis and Carolyn T Thorpe. 2021. Scale development: Theory and applications. Sage publications.

[27] Brian S Everitt. 1975. Multivariate analysis: The need for data, and other problems. The British Journal of Psychiatry 126, 3 (1975), 237–240.

[28] R Michael Furr. 2021. Psychometrics: an introduction. SAGE publications.

[29] Michael A Goodrich, Alan C Schultz, et al. 2008. Human–robot interaction: a survey. Foundations and Trends® in Human–Computer Interaction 1, 3 (2008), 203–275.

[30] Heather M Gray, Kurt Gray, and Daniel M Wegner. 2007. Dimensions of mind perception. science 315, 5812 (2007), 619–619.

[31] Kathy E Green, Catherine G Frantom, et al. 2002. Survey development and validation with the Rasch model. In International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC. 14–17.

[32] Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology 31, 4 (2016), 337–350.

[33] Edward Guadagnoli and Wayne F Velicer. 1988. Relation of sample size to the stability of component patterns. Psychological bulletin 103, 2 (1988), 265.

[34] Gregory R Hancock and Ralph O Mueller. 2001. Rethinking construct reliability within latent variable systems. Structural equation modeling: Present and future 195 (2001), 216.

[35] Donna Harrington. 2009. Confirmatory factor analysis. Oxford university press.

[36] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology. Vol. 52. Elsevier, 139–183.

[37] Larry Hatcher and Norm O'Rourke. 2013. A step-by-step approach to using SAS for factor analysis and structural equation modeling. Sas Institute.

[38] Ville Heilala, Riitta Kelly, Mirka Saarela, Päivikki Jääskelä, and Tommi Kärkkäinen. 2023. The Finnish version of the affinity for technology interaction (ATI) scale: psychometric properties and an examination of gender differences. International Journal of Human–Computer Interaction 39, 4 (2023), 874–892.

[39] Chin-Chang Ho and Karl F MacDorman. 2010. Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. Computers in Human Behavior 26, 6 (2010), 1508–1518.

[40] Li-tze Hu and Peter M Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural equation modeling: a multidisciplinary journal 6, 1 (1999), 1–55.

[41] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. International journal of cognitive ergonomics 4, 1 (2000), 53–71.

[42] Alexandra D Kaplan, Tracy L Sanders, and Peter A Hancock. 2021. Likert or not? How using Likert rather than biposlar ratings reveal individual difference scores using the Godspeed scales. International Journal of Social Robotics 13, 7 (2021), 1553–1562.

[43] Paul Kline. 2013. Handbook of psychological testing. Routledge.

[44] Rex B Kline. 2023. Principles and practice of structural equation modeling. Guilford publications.

[45] Mika Koverola, Anton Kunnari, Jukka Sundvall, and Michael Laakasuo. 2022. General attitudes towards robots scale (GAToRS): A new instrument for social surveys. International Journal of Social Robotics 14, 7 (2022), 1559–1581.

[46] Lisa Schurer Lambert and Daniel A Newman. 2023. Construct development and validation in three practical steps: Recommendations for reviewers, editors, and authors. Organizational Research Methods 26, 4 (2023), 574–607.

[47] Jon Landeta. 2006. Current validity of the Delphi method in social sciences. Technological forecasting and social change 73, 5 (2006), 467–482.

[48] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. Human factors 46, 1 (2004), 50–80.

[49] Kwan Min Lee, Namkee Park, and Hayeon Song. 2005. Can a robot be perceived as a developing creature? Effects of a robot's long-term cognitive developments on its social presence and people's social responses toward it. Human communication research 31, 4 (2005), 538–563.

[50] Benedikt Leichtmann, Verena Nitsch, and Martina Mara. 2022. Crisis ahead? Why human-robot interaction user studies may have replicability problems and directions for improvement. Frontiers in Robotics and AI 9 (2022), 838116.

[51] John M Linacre, MH Stone, J William, P Fisher, and L Tesio. 2002. Rasch Measurement. Rasch Measurement Transactions 16 (2002).

[52] Harold A Linstone, Murray Turoff, et al. 1975. The delphi method. Addison-Wesley Reading, MA.

[53] Robert C MacCallum, Keith F Widaman, Shaobo Zhang, and Sehee Hong. 1999. Sample size in factor analysis. Psychological methods 4, 1 (1999), 84.

[54] Bertram Malle. 2019. How many dimensions of mind perception really are there?. In CogSci. 2268–2274.

[55] Bertram F Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In Trust in human-robot interaction. Elsevier, 3–25.

[56] D Betsy McCoach, Robert K Gable, and John P Madura. 2013. Instrument development in the affective domain. Vol. 10. Springer.

[57] S. Messick. 1989. In R. L. Linn (Ed.) Educational Measurement. American Council on Education and National Council on Measurement in Education, Washington, D.C., 12–103.

[58] Jennifer L Monahan. 1998. I don't know it but I like you: The influence of nonconscious affect on person perception. Human Communication Research 24, 4 (1998), 480–500.

[59] Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kennsuke Kato. 2004. Psychology in human-robot communication: An attempt through investigation of negative attitudes and anxiety toward robots. In RO-MAN 2004. 13th IEEE international workshop on robot and human interactive communication (IEEE catalog No. 04TH8759). IEEE, 35–40.

[60] Jum C Nunnally. 1978. Psychometric Theory. McGraw-Hill.

[61] Steven J Osterlind. 2006. Modern measurement: Theory, principles, and applications of mental appraisal. Pearson/Merrill Prentice Hall Upper Saddle River, NJ.

[62] Aaron Powers and Sara Kiesler. 2006. The advisor robot: tracing people's mental model from a robot's physical attributes. In Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction. 218–225.

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

[64] William Revelle. 2022. An introduction to psychometric theory with applications in R.
[65] William Revelle and Richard E Zinbarg. 2009. Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. Psychometrika 74 (2009), 145–154.
[66] Mark J Schervish. 1996. P values: what they are and what they are not. The American Statistician 50, 3 (1996), 203–206.
[67] John A Schinka, Wayne F Velicer, and Irving B Weiner. 2013. Handbook of psychology: Research methods in psychology, Vol. 2. John Wiley & Sons, Inc.
[68] Mariah Schrum, Muyleng Ghuy, Erin Hedlund-Botti, Manisha Natarajan, Michael Johnson, and Matthew Gombolay. 2023. Concerning trends in likert scale usage in human-robot interaction: Towards improving best practices. ACM Transactions on Human-Robot Interaction 12, 3 (2023), 1–32.
[69] Klaas Sijtsma. 2009. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 74 (2009), 107–120.
[70] Valmi D Sousa and Wilaiporn Rojjanasrirat. 2011. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. Journal of evaluation in clinical practice 17, 2 (2011), 268–274.
[71] J. Gregory Trafton, Chelsea R. Frazier, Kevin Zish, Branden J. Bio, and J. Malcolm McCurry. 2023. The Perception of Agency: Scale Reduction and Construct Validity*. In 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). 936–942. https://doi.org/10.1109/RO-MAN57019.2023.10309544
[72] J Gregory Trafton, Paula Raymond, and Sangeet Khemlani. 2021. The power of theory. ACM Transactions on Human-Robot Interaction (THRI) 10, 1 (2021), 1–3.
[73] Daniel Ullman and Bertram F Malle. 2018. What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In Companion of the 2018 acm/ieee international conference on human-robot interaction. 263–264.
[74] Bertie Vidgen and Taha Yasseri. 2016. P-values: misunderstood and misused. Frontiers in Physics 4 (2016), 6.
[75] Rebecca M Warner and David B Sugarman. 1986. Attributions of personality based on physical appearance, speech, and handwriting. Journal of personality and social psychology 50, 4 (1986), 792.
[76] Ronald L Wasserstein and Nicole A Lazar. 2016. The ASA statement on p-values: context, process, and purpose. , 129–133 pages.
[77] Kara Weisman, Carol S Dweck, and Ellen M Markman. 2017. Rethinking people's conceptions of mental life. Proceedings of the National Academy of Sciences 114, 43 (2017), 11374–11379.
[78] Stefanie Wind and Cheng Hua. 2021. Rasch measurement theory analysis in R: Illustrations and practical guidance for researchers and practitioners. Bookdown. org,[Epub] (2021).
[79] Benjamin D Wright and Mark H Stone. 1979. Best test design. (1979).
[80] 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 (2012), 235–248.
[81] Matthias Ziegler and Dirk Hagemann. 2015. Testing the unidimensionality of items.
[82] Megan Zimmerman, Shelly Bagchi, Jeremy Marvel, and Vinh Nguyen. 2022. An analysis of metrics and methods in research from human-robot interaction conferences, 2015–2021. In 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 644–648.
[83] Richard E Zinbarg, William Revelle, Iftah Yovel, and Wen Li. 2005. Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. psychometrika 70 (2005), 123–133.

A APPENDIX: PRINTER-FRIENDLY VERSION OF THE GUIDELINE

Table A1. Printer-Friendly Guideline – Choosing the Perfect Scale

Stage	Question	Check
Item Development	1. Is the construct clearly defined?	☐
	2. Is the item generation process discussed (e.g., via a literature review, the Delphi method, or crowd-sourcing)?	☐
	3. Do the final items capture the construct as it has been defined in the paper?	☐
Scale Development	4. Full initial set of items reported?	☐
	5. Does the sample size meet the 10 (participants) : 1 (initial number of items) criteria?	☐
	6. Did scale developers report using EFA, PCA, or Rasch to determine item:factor relationship?	☐
	7. Factor extraction method discussed?	☐
	8. Factor loadings (EFA/CFA) or item fits (Rasch) for all items provided?	☐
	9. Is the item removal process (e.g., using infit/outfit, factor loading minimum values, or cross-loading values) described?	☐
	10. Complete list of items in the final version of scale reported?	☐
Scale Evaluation	11. Test for factor structure reported (e.g., second EFA, CFA, DIF, test of unidimensionality if using Rasch, or similar)?	☐
	12. Reliability reported (e.g., Cronbach's α, McDonald's ω_h or ω_t, Tarkkhonen's Rho)?	☐
	13. Was a test of validity reported (e.g., predictive, concurrent, convergent, discriminant)?	☐