Choosing the "perfect scale": a scale development and selection primer for non-experts

Journal:	*Transactions on Human-Robot Interaction*
Manuscript ID	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Saad, Laura; US Naval Research Laboratory, Roesler, Eileen; George Mason University Phillips, Elizabeth; George Mason University, Psychology Trafton, J. Gregory; US Naval Research Lab,
Keywords:	scale development, psychometrics, human-robot interaction, questionnaires

Choosing the "perfect scale": a scale development and selection primer for non-experts

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

Scales are commonly employed in HRI research yet 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 that aims to empower researchers without scale development expertise with the tools to assess scale quality efficiently. The guideline provides high-level questions and examples to help the reader make confident evaluations of pre-existing scales. 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 can 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 scale development and selection primer for non-experts. 1, 1 (January 2024), 16 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, self-report measures (or scales), 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 [59] and of those 61% (889 papers) included scales.

The HRI and RO-MAN community, which is inherently multidisciplinary, is composed of roboticists, computer scientists, psychologists, cognitive scientists, designers, linguists, and engineers, among others [24]. Many of these disciplines do not provide direct training in psychometrics, the scientific study of testing, measurement, and assessment in the behavioral sciences. This makes it likely that researchers incorporate scales as dependent measures in their experiments without

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

ACM XXXX-XXXX/2024/1-ART
https://doi.org/XXXXXXX.XXXXXXX

expertise in determining whether that scale is appropriate for the research goals and whether the scale has even 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 and development, as well as selection and evaluation. This is particularly relevant to those aiming to advance their research within the HRI community, but also for those who value reproducible research [36]. Developing the ability to critically analyze pre-existing scales will become a necessary tool in the proverbial toolbox if the HRI community wishes to ensure continued impact in both academic contexts and society more broadly.

Developing a well-designed scale not only takes a substantial amount of time but also requires both art and science [6, 20]. The field of psychometrics, originating with William Cattell's *Mental Tests and Measurements* [15], 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). The ease of implementation of a scale into a project combined with a lack of expertise on appropriate 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"[1]) they measure or to a broader psychological theory can be problematic as it limits the interpretation of the outcomes given by the scale [52]. Additionally, without some knowledge regarding the best practices in psychometrics, simple mistakes and assumptions can propagate easily leading to work that is potentially uninterpretable. Avoiding this immense waste of resources should be an appealing goal to the entire HRI community and provides partial motivation for this paper.

The aim of this paper 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 guide on the basics of scale development. The guideline aims to provide the reader with high-level conceptual information that can help them determine the quality of pre-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. In part five, the paper concludes with some suggestions for future work in the domain and some additional resources for those interested in a more in-depth understanding of scale development.

2 GUIDELINES FOR CHOOSING THE "PERFECT" SCALE

The term "scale" refers to any self-report or 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 pre-existing scale, or 3) creating a customized scale. Developing a new scale might be required in cases where the construct of interest has been difficult to measure in the past or particularly in cases where the construct has never been measured before. Therefore, there are cases where developing a new scale might be necessary.

[1]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.

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 construct of interest. This paper aims to help the reader select the appropriate scale and evaluate its validity as a measure. Depending on the construct of interest, it is possible there will be multiple scales to choose from and so 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. These guidelines consist of eight high-level questions that should encourage the reader to engage in the critical analysis of a scale before incorporating it into their study. These questions were synthesized from other sources [6, 20] and the authors' experience analyzing and developing scales.

The scale development process can be separated into three stages: 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 that can allow the reader to determine whether the scale has met these 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 a focal 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 [16]. This section lists two main questions the reader should ask of the scale they are evaluating to ensure that the item development process was completed adequately and the construct the scale is intended to measure is appropriately defined.

2.1.1 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 that the project aims to measure or investigate. In psychometrics, this is referred to as identifying the construct of interest. A construct refers to the concept or unobserved behavior that is the target of the study [6]. Measurement scales are typically developed to measure a latent construct that can be inferred from participant's responses to scale items [23, 45]. The spectrum of possible constructs that can be measured is expansive and, for example, can range from performance [29], to perceptions [40], to preferences [10]. Regardless, a clear definition of the construct is critical.

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 [6]. Ideally, a precise definition that is agreed upon will already exist in the literature in which case the theory-driven approach is the best avenue forward. For example, Carpinella et al. [14] used a theory-driven approach to determining the construct of social perception of robots as the perception of robots' warmth and competence. 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. It is also appropriate, in the case where no agreed-upon theoretical framework exists, 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. [25] 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 mind perception researchers have also used this approach to examine the latent dimensions of mind perception [39, 55]. The data-driven approach is useful as it allows the data to "speak for itself" [55]. 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 explanation 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?" The purpose of evaluating the construct definition is to enable the reader with the appropriate 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.

2.1.2 Question 2: Do the items reported seem to capture all aspects of the definition reported? 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 adequately critique the items that were included in the scale and therefore cannot ensure that the measure is 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 different constructs. To determine whether this is the case, it can be helpful for the reader to refer back to the definition of the construct to see if it includes an explanation of behaviors that the proposed construct does not encompass. For example, in their measure of trust, Malle and Ullman [40] stated definition was multidimensional and their study identified performance and morality as the two main factors (or dimensions) that underlie the experience 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 [40].

For non-experts, it can be easy 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. It is important that the reader considers this during the item review process. If the items do not match the stated construct then the reader can assume that the scale is not the best measure of the construct of interest and may not be useful to include in their project.

In these cases, it may be useful to consider incorporating the Delphi method for item development [38]. 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) [34]. Though not always possible, we recommend consulting experts in the field during the item development process, especially in cases where the individual developing the scale is not a subject matter expert in the construct of interest.

A construct can be unidimensional or multi-dimensional. A unidimensional scale measures the construct along a single range from low to high. For example, the perception of agency scale [51] is a unidimensional scale. More common are multi-dimensional scales like trust [16, 40, 53], negative attitudes towards robots [43], 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. It can be particularly tricky to determine which items should be included in a scale creation – the definition of the construct is critical for this stage; if items do not relate to the definition, additional dimensions may appear that are unrelated to the overall construct.

Lastly, the reader should also ensure that 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 [22]. 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. So including items that are grammatically redundant can improve the reliability while simultaneously decreasing the overall usefulness of the scale as a measure of the construct of interest because the scale may be missing items needed to capture all the dimensions of the construct.

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

2.1.3 BONUS: Is there any mention of item generation in the paper? 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 majority of the dimension (or dimensions) that make up the construct of interest.

2.2 Stage 2: Scale Development

Boateng et al. [6] 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 [56] 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, i.e., a process for fitting data to a model [1, 56].

Though there are many different methods for developing a scale, there are some components of the process that are consistent across methods. 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 study as well as the type of development method the paper uses. 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. Lastly, the reader should 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. The rest of this section provides specific details for the reader to investigate whether the development of the scale was adequate.

2.2.1 Question 3: Does the test sample size meet the 10:1 minimum criteria? The initial sample of a new scale should include all items, not just the "good" ones. The sample size should follow the 10:1 criteria [44]. This refers to 10 participants per item on the scale. However, even larger samples are usually better [26] since they ensure lower measurement error and more stable factor loadings.

Take home: Sample sizes for validation studies should follow the 10:1 (people to items) rule though more participants is considered a positive feature.

2.2.2 Question 4: Do the authors adequately analyze and report on the relationship of items to dimension within the construct of interest? 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 psychometrics, 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 principal components analysis (PCA), exploratory factor analysis (EFA), 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 some heuristics that are frequently used in the field. (The inclined reader can learn more about factor analysis and other methods of investigating the relationship between items and dimensions in the following resources: [23, 33, 45, 47, 48, 56].) What the reader should ensure is that at the very least the study should include some description of how the authors determined the number of dimensions (which we refer to as the scale development method and can include PCA, factor analysis, or Rasch) and how many dimensions the final scale includes.

After determining the number of dimensions, the paper should report the relationship between the items and the construct of interest. For 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 [44]. 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. Rasch analysis uses outfit and infit measures [7, 57]. Generally, Rasch items show poor fitting items when an outfit is higher than 1.5 [37].

Take home: There are many methods that can be used to determine the underlying factor structure of the construct of interest. There are three questions the reader can ask to determine whether the item to factor to construct relationship was adequately discussed:

1) was the scale development method described, 2) was there a description of how the number of factors was determined, and 3) was there some quantitative description of the relationship between each item and any constructs.

2.2.3 Question 5: Do the authors elaborate on how items were removed? 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 can be removed based on low factor loadings or high cross loadings of one item across multiple factors. Additionally, the Rasch model includes fit statistics such as infit or outfit that can be used to determine item fit to the construct. Again, the details are not necessarily important for the non-expert; what is important is determining whether or not the scale had appropriate and consistent criteria for this process.

Items can be removed not only due to lack of fit within the domain but also due to redundancy with other items [22]. 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, potentially unrelated 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.

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. It is important that authors report their processes for item reduction.

2.3 Stage 3: Scale Evaluation

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

2.3.1 Question 6: Is the factor structure the same as when the scale was created? 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 [58]. 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) [11, 12, 28], though others can also be used. Each fit statistic has a heuristic

value that the CFA should be under (or over). A CFA should thus report some measure(s) of fit and what the acceptable range is.

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) [8, 56]. 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 researcher should perform another CFA on a different group of participants to examine 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.

2.3.2 Question 7: Do the authors report a measure of reliability for the entire scale? Reliability refers to the principle that a measurement produces similar results under similar conditions and is related to one of the core components of science: replicability. In addition, reliability is a starting place for establishing scale validity, as a measure cannot be more valid than it is reliable. 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 [20, 49]. 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 [17].

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 [18, 20, 27, 50, 60] 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 [44]. Interestingly, a closer inspection of the original text shows that this is in fact a misrepresentation (as has been previously noted in [20]). Nunnally [44] 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, particularly in cases where the results will inform decisions that impact society as is the case with some HRI research. As a result, we suggest ≥ 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.

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 researcher'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 if it is less than ideal) 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, regardless of metric, is a value ≥ 0.80 though papers should report reliability even if it is below this threshold.

2.3.3 Question 8: Are there explicit tests of validity reported? Reliability is not the same as validation 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 [21, 41, 47]. The concept of validity can be split into sub-components such as criterion and construct validity [6, 21]. Criterion validity refers to the degree to which there is a relationship between behavior on the current scale and behavior 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 and concurrent validity measures the degree to which the performance on the current scale relates to performance on a criterion (gold standard) measurement [21]. 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, 9]. Construct validity can be measured in many ways [13, 19] 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 and 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 you are developing or validating.

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 not assume it is a valid measure of the construct.

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 the previous section.

The first stage, item development, consists of determining whether the scale has a construct that has been precisely defined, either by starting with a precise theory from the literature or by using a bottom-up approach where the definition of the construct comes from the data structure that is revealed after pilot testing. Additionally, this stage requires the reader to 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 the desired construct. This 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 ensure that the pilot sample is large enough to conduct the appropriate analyses. The ideal sample size is at least 10 participants for every item (e.g., 120 participants for a 12-item scale). If the scale meets that minimum requirement, then the reader can proceed to determine whether and how the items capture the underlying structure of the construct of interest. Though there are many different and accepted methods for this process, the reader should look for at least one method that details some explanation of 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 a good validation report will have some discussion on this process, particularly regarding whichever threshold or inclusion/exclusion criteria was used.

The third and final stage, scale evaluation, consists of reliability and validity checks. First, the reader should look for a check regarding the consistency of the reported factor structure. Second, the scale should have some acceptable measure of reliability. Ideally, the authors will have computed ω_t or ω_h for the scale, depending on the dimensionality, in addition to the typical Cronbach's coefficient α. Regardless of the measure used, we suggest that your reliability should be ≥ 0.80. Lastly, a test of validity should be included in the report.

Armed with this information, the reader should now feel more confident in critically analyzing pre-existing scales and their corresponding validation reports. To guarantee this confidence, we next briefly turn to two examples from the HRI literature and walk through the guidelines 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 guidelines presented in this paper to evaluate two scales that are frequently used in HRI – the Godspeed Questionnaire [5] and the Robotic Social Attribute Scale [14]. 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 with the goal of developing a tool that can be used 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 [14]. 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 [35, 42, 46, 54] 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. 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

Table 1. Applying the Guideline to Two HRI Scales

Stage	Question	Godspeed	RoSAS
Item Development	1. Construct defined?	✓	✓
Item Development	2. Do items capture construct?	×	✓
Item Development	Items are unambiguous and clear?	×	✓
Item Development	Items are related to construct?	×	✓
Item Development	BONUS: Item generation process discussed?	✓	✓
Scale Development	3. Person:Items 10:1?	×	✓
Scale Development	4. Item/factor relationship discussed?	×	✓
Scale Development	Is scale development method described?	×	✓
Scale Development	Description of how number of factors was determined?	×	✓
Scale Development	Described type of rotation used (if using EFA)?	×	✓
Scale Development	5. Description of item removal process?	×	✓
Scale Evaluation	6. Test for factor structure?	×	×
Scale Evaluation	7. Reliability reported?	✓	✓
Scale Evaluation	8. Validity reported?	×	✓

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 two of the criteria: adequate construct definition and reliability test results reported. However, details about how reliability was conducted were actually reported in other reports [3, 4] which were not scale development papers, and only reliability (α) was reported. Since the anthropomorphism items within Godspeed are modified versions of the originals from the "humanlikeness" scale reported in Powers and Kiesler [46] it can be argued that the guideline should be applied directly to that scale. In doing so, we see that additional criteria were met, specifically that the item generation process is much clearer. However, importantly, the reader can make a better assessment of the adequacy of the scale's development and may feel more confident in using the Powers and Kiesler [46] version of the items compared to those incorporated in the Godspeed scale [5].

Additionally, although the specific type of custom scale approach that Bartneck et al. [5] used to develop the Godspeed scale is not ideal it does not mean that it is inappropriate 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 on how well the items fit together to measure the hypothesized dimensions underlying 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 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 [14, 30, 32]). 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.

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. Do items capture construct?	×	✓
Item Development	Items are unambiguous and clear?	×	✓
Item Development	Items are related to construct?	×	✓
Item Development	BONUS: Item generation process discussed?	×	✓
Scale Development	3. Person:Items 10:1?	×	×
Scale Development	4. Item/factor relationship discussed?	×	✓
Scale Development	Is scale method described?	×	✓
Scale Development	Description of how number of factors was determined?	×	×
Scale Development	Described type of rotation used (if using EFA)?	×	×
Scale Development	5. Description of item removal process?	×	×
Scale Evaluation	6. Test for factor structure?	×	×
Scale Evaluation	7. Reliability reported?	✓	✓
Scale Evaluation	8. 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. The sample size for the pilot studies was appropriate with at least 200 participants per study which is greater than 10 participants for each of the 18 items that were included in the final version of the scale. The only missing component to the RoSAS development process was there was no report of a factor structure test (e.g., no confirmatory factor analysis) making it unclear whether the factor structure remains consistent across different samples. 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 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 criteria for reliability and validation as per the criteria in the guideline outlined in this paper. 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 pre-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 psychometrically validated as described in this article. If it was done correctly, the scale should have a strong majority of checks using our approach. However, sometimes a needed scale may be too niche and 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[2]. 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.

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" [31], etc. are all acceptable.

Changing the response range (e.g., limits of a Likert scale) of an existing scale is not recommended. In some cases, it can change the meaning of the scale (e.g., converting a scale with an even number of response categories with no neutral into a scale with a midpoint can change how people would respond to it). In some analysis methods (e.g., Rasch), the exact response range is critical to generating acceptable data.

If a researcher does decide to generate a custom scale, we recommend that the researcher be explicit that the scale is a custom scale, generating 4-6 items that the researcher believes best capture the latent variable of interest, 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, for example, 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 potential limitation of the measure in the paper and suggest (or conduct) a follow-up study which includes a longer, validated scale to which the single item could be compared. This is particularly important in cases where there are no other validated scales that measure the construct.

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 are better equipped to critically analyze and implement pre-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

[2]Subscales refer to complete sets of items that load onto one factor in a pre-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 [14]. Including only a subscale in a study is completely acceptable.

appropriately is not a simple process. We hope that our guideline has provided the information necessary to determine whether that scale was 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 authors only reported $\alpha > 0.70$) there are additional steps to 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 can 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 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. Those interested in learning more about the scale development process can begin with these resources [6, 8, 20, 37, 48, 56, 57].

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

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

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

[10] John Brooke. 1996. Sus: a “quick and dirty’usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.

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

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

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

[14] 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.

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

[16] 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.

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

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

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

[20] 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.

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

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

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

[24] 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.

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

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

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] Paul Kline. 2013. *Handbook of psychological testing*. Routledge.

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

[35] 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.

[36] 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.

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

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

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

[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] Jum C Nunnally. 1978. *Psychometric Theory.* McGraw-Hill.
[45] Steven J Osterlind. 2006. *Modern measurement: Theory, principles, and applications of mental appraisal.* Pearson/Merrill Prentice Hall Upper Saddle River, NJ.
[46] 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.
[47] Tenko Raykov and George A Marcoulides. 2011. *Introduction to psychometric theory.* Routledge.
[48] William Revelle. 2022. An introduction to psychometric theory with applications in R.
[49] William Revelle and Richard E Zinbarg. 2009. Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika* 74 (2009), 145–154.
[50] Klaas Sijtsma. 2009. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74 (2009), 107–120.
[51] 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
[52] 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.
[53] 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.
[54] 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.
[55] 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.
[56] Stefanie Wind and Cheng Hua. 2021. Rasch measurement theory analysis in R: Illustrations and practical guidance for researchers and practitioners. *Bookdown. org,[Epub]* (2021).
[57] Benjamin D Wright and Mark H Stone. 1979. Best test design. (1979).
[58] Matthias Ziegler and Dirk Hagemann. 2015. Testing the unidimensionality of items.
[59] 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.
[60] 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.