Choosing the "Perfect" Scale: A Primer to Evaluate Existing Scales in HRI

LAURA SAAD, US Naval Research Laboratory, Washington, District of Columbia, USA
EILEEN ROESLER and ELIZABETH PHILLIPS, George Mason University, Fairfax, Virginia, USA
J. GREGORY TRAFTON, US Naval Research Laboratory, Washington, District of Columbia, USA

Scales are commonly employed in Human–Robot Interaction (HRI) research, yet due to its multidisciplinary nature, many in this community lack direct training in psychometrics. This poses challenges for appropriate scale selection, accurate assessments of reliability and validity, and use. We provide a tutorial to empower researchers without scale development expertise to assess scale quality efficiently. We detail a guideline that 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, whereas Godspeed warrants further investigation before it should be used in HRI contexts. The article 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 Phillips, and J. Gregory Trafton. 2026. Choosing the "Perfect" Scale: A Primer to Evaluate Existing Scales in HRI. *ACM Trans. Hum.-Robot Interact.* 15, 2, Article 42 (January 2026), 30 pages.
https://doi.org/10.1145/3772066

1 Navigating the Challenges of Scale Development and Selection

Measurement is a fundamental aspect of scientific research. Scientific discovery depends upon accurate, reliable, and valid measures and developing such measures requires a principled approach [36]. In the field of **Human–Robot Interaction (HRI)** research, scales (also known as rating scales or questionnaires), are a key methodological tool. They offer insights into user perspectives and experiences, and provide critical data that informs the design and evaluation of robotic systems. An analysis of the publications from 2015 to 2021 across two major HRI venues, the ACM/IEEE

This work was supported in part by ONR to J. G. Trafton.
Authors' Contact Information: Laura Saad (corresponding author), US Naval Research Laboratory, Washington, District of Columbia, USA; e-mail: laura.s.saad.ctr@us.navy.mil; Eileen Roesler, George Mason University, Fairfax, Virginia, USA; e-mail: eroesle@gmu.edu; Elizabeth Phillips, George Mason University, Fairfax, Virginia, USA; e-mail: ephill3@gmu.edu; J. Gregory Trafton, US Naval Research Laboratory, Washington, District of Columbia, USA; e-mail: greg.j.trafton.civ@us.navy.mil.

ACM 2573-9522/2026/1-ART42
https://doi.org/10.1145/3772066

International Conference on HRI and the IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), found that of the 1,464 published papers [126], 61% (889 papers) included scales. This widespread use underscores the importance of scales in advancing research in HRI and social robotics more broadly.

The social robotics community is inherently interdisciplinary, composed of roboticists, computer scientists, psychologists, cognitive scientists, designers, linguists, and engineers, among others [44]. While this diversity has many advantages, it also means that many researchers in the field do not have formal training in psychometric theory,[1] the branch of psychology dedicated to the scientific study of testing, measurement, and assessment of psychological constructs [97]. Psychometric theory offers the tools needed to evaluate whether a scale accurately captures the constructs it claims to measure and whether it does so reliably. Those looking to learn more about psychometric theory will quickly find that the field 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 (see Cattell's *Mental Tests and Measurements* [24] for a foundational reference). Gaining expertise in this field is not a simple task (and will not be accomplished by simply reading one paper).

Yet, even without much experience in psychometric theory, it is quite easy to implement a scale into a research 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 development, selection, and use [105]. Without some knowledge regarding the best practices in psychometric theory, simple mistakes and assumptions can propagate easily, leading to work that may be difficult to replicate and potentially uninterpretable. These skills are particularly relevant to not only those aiming to advance their research within the HRI community, but also to those who value reproducible research more broadly [74]. Developing the ability to critically analyze existing scales will become a necessary tool in the proverbial toolbox if the HRI community aim to achieve more valid and reproducible results which will in turn ensure continued positive impact in both academic and applied contexts.

Therefore, to help build these critical skills, we offer this tutorial aimed at equipping members of the HRI community, particularly those without expertise in scale development, with a practical tool to critically assess whether a scale has been adequately developed and validated. The current section (Section 1) provides motivation for the development of this tool. Section 2 introduces a step-by-step guideline on the basics of scale development, outlining the essential steps a scale must undergo to be considered well-developed. The goal of the guideline is to equip readers with the knowledge needed to critically assess the quality of existing scales. Section 3 applies the guideline to evaluate two frequently used and cited scales in HRI, Godspeed [10] and the **Robotic Social Attributes Scale (RoSAS)** [23]. Section 4 provides advice for those interested in using customized scales in their research. Readers should note that this tutorial is not intended as a guide for developing a new scale. Those interested in learning more on this topic can refer to Section 5, which concludes the article with suggestions for future work and provides additional resources to support 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 article.

2 Guidelines for Choosing the "Perfect" Scale

The term "scale" refers to any instrument that measures an attitude, attribute, or other latent construct that is not directly observable [38].[2] 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 [89], between- or within-subjects [90]), 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—a custom scale is any scale that has not been validated (see Section 4 for more details on custom scales).

Developing a new scale might be required in several situations. First, this can occur when the construct of interest has been difficult to measure in the past (e.g., the construct is complex, vague, or does not fit neatly within a previously proposed definition). Second, a new scale might be needed when the construct has never been measured before, or when an existing scale is available but has methodological flaws. Finally, scale development might be necessary when the theory underlying the construct has evolved and the scale needs to be updated accordingly.

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 article (though see Section 5 for some suggested readings). Therefore, a reasonable first step is to determine whether or not a scale already exists that adequately measures the attitude or attribute of interest.

This tutorial aims to help the reader select an appropriate scale, evaluate whether it was developed adequately, and determine its validity as a measure. In many cases, multiple scales may be available, and selecting the "perfect" scale depends not only on how well it fits with the research aims but also on its measurement abilities. A "perfect" scale is one that both measures the intended construct and has been rigorously developed and validated. Only when previously validated scales are not available (or the existing scales have not been adequately developed and validated) should the reader proceed to develop a custom scale. Section 4 provides information and advice for those developing custom scales.

The following subsections detail a guideline for choosing the "perfect" scale from existing scales in HRI. The guideline consists of 13 high-level questions designed to help the reader evaluate whether a candidate scale is a good fit for their research. These questions were synthesized from other sources [13, 34] and the authors' (of this article) experience analyzing and developing scales.

The authors would like to note that many of the guideline items include recommendations for minimum acceptable criteria for specific metrics used in the scale development process (e.g., Cronbach's alpha, **Comparative Fit Index (CFI)**, and RMSEA). 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 article. In the likely event that a chosen scale meets most, but not all, of the guideline criteria, it may still be suitable for use. In such cases, the reader should carefully examine which aspects of the criteria the scale does not meet and consider the implications of those gaps in the context of their research goals. A scale may seem to align well with a reader's intended construct based on its content or theoretical framing, but, if it lacks evidence of rigorous development and validation, its usefulness may be limited. In these situations, readers are encouraged to consider alternative scales that, while not an exact conceptual match,

[2]The terms construct, domain, and latent variable are frequently used interchangeably in the psychometric literature and all refer to the same thing—the unobservable attitude or attribute that is being measured [40].

may offer stronger psychometric support and ultimately better serve the goals of their research.[3] Importantly, if the reader's chosen scale fails to meet the majority of the guideline criteria, this raises substantial concerns about its measurement abilities and warrants reconsideration of its use.

We also acknowledge that different scale developers will use slightly different heuristics across studies and development methods. This is perfectly acceptable in scientific research so long as an adequate motivation or rationale is provided. Scale developers 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 statistical "significance" in science more broadly. Though this threshold is far from perfect [49, 103, 116, 118], it can be used as a guard rail to aid researchers in the interpretation of results [33]. An experienced researcher can recognize that the 0.05 threshold is arbitrary and that a p-value of 0.054 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, an ω value of 0.68 is not meaningfully different from an ω value of 0.72 (ω here refers to McDonald's ω. See [34, 57] for more details.); both values suggest that there are reasonably high levels of internal reliability in the measure. When possible, we recommend that the reader consider approaching the heuristics outlined in this tutorial with similar nuance. However, we recommend using these criteria as a first step for understanding basic scale development criteria, given that the target audience for this tutorial are readers without expertise in psychometric theory.

The guideline is separated into the three main stages of the scale development process: item development, scale development, and scale evaluation. At each stage, this article identifies the minimum requirements a "good" scale must meet and poses specific questions that 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 construct (i.e., attitude or attribute) 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 [26]. 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 scale construct 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) attitude, cognition, or attribute that is the target of the study [14, 67]. Unobserved (or latent) 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 participants' responses to scale items [43, 93].

[3]Relatedly, if there are few scales available that measure the construct of interest and these scales have not been adequately developed or validated, it may still be acceptable to use one of them in a research project. However, we recommend the researcher critically evaluate the scale (e.g., with the methods detailed below) and share their findings to help advance understanding of its measurement abilities.

An example of a latent construct relevant to the HRI might be perceived agency, which is defined as "a characteristic of an entity whose actions are assumed by an outside observer to be driven primarily by its internal thoughts and feelings and less by the external environment" [112, p. 3]. 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, making it a clear example of a latent construct. The spectrum of possible latent constructs that can be measured is expansive and, for example, can range from constructs related to performance [55], perceptions [80], or preferences [18].

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. An example of a good definition is one by Malle and Ullman [80] who define trust as "a 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. 12). This definition is good because it clearly provides hypothesized factors that are thought to be contained within the construct (i.e., performance trust and moral trust) as well as key aspects of the context in which trust is thought to occur (i.e., cases where an individual in the dyad is vulnerable). Additionally, a good definition should not only state what the construct is, but ideally also what it is not [70], although this is not as common an occurrence (though see [81] for a nice example of a literature review that distinguishes their proposed construct from other related constructs in the field).

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 [83]. Ideally, a precise definition that is agreed upon will already exist in the literature. In this case, the theory-driven approach is the best avenue forward. 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. For example, Malle and Ullman [80] synthesized decades of research to develop their definition of trust. Alternatively, in cases where there is no theoretical consensus of the construct, scale developers with expertise in the area can choose one of the debated theories, motivate this decision, and develop items based on this definition. However, without agreement in the literature 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 are thought to be related to the construct of interest. For example, in their efforts to investigate the dimensions underlying the mind perception of others, Weisman et al. [119] included a range of items that captured many different potential underlying dimensions and then evaluated those dimensions *post hoc* to develop their body–heart–mind framework. Other researchers have also used a data-driven approach to examine the latent dimensions of mind perception [79, 82]. The data-driven approach is useful as it allows the data to "speak for itself" [119]. 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*.

Recognizing whether a scale was developed using a theory- or data-driven approach is important for determining whether the construct is clearly defined (i.e., applying guideline question 1). In the case of the theory-driven approach, the definition may be clearly linked to existing theoretical frameworks and may be stated toward the beginning of the scale development paper. If a scale developer used a data-driven approach, then the definition will likely arrive toward the end of the paper once the analysis has revealed the underlying structure of the construct. Regardless, both

approaches should lead to the same destination: a clear definition of the construct that is explicitly stated somewhere in the paper.

In some cases, one construct will have different scales that state different, and sometimes competing, definitions of a proposed construct. For example, Malle and Ullman [80] and Lee and See [72] explicitly defined trust in terms of vulnerability assumed by one individual in a dyad, while Yagoda and Gillan [124] defined the construct in terms of performance, function, and semantics. Different definitions highlight alternative views on complex constructs and promote growth in the field by encouraging empirical comparisons. For example, in the context of the aforementioned competing trust definitions, a researcher could design an experiment to test whether vulnerability is a necessary component in developing trust with a robot. These types of investigations then strengthen the existing understanding of the construct either by affirming one definition over another, or by pushing researchers to consider additional factors that have been previously overlooked.

Additionally, 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 more broadly will likely not want to use a measure that has defined trust in other contexts (e.g., Charalambous' scale for use in industrial contexts [26]). A good definition can also help researchers more precisely determine what it is they are interested in investigating. For example, a researcher may initially believe they are interested in trust but after reading through the stated construct definitions they may realize that their true interest is in acceptance or intention to use the robot. This determination is the first, informal, step in the process of determining construct validity, or the degree to which a measurement tool accurately measures what its intended to measure (see guideline question 3 for more details on this topic).

Importantly, we recommend the reader consider both their research goals and the quality of the scale development process together when making a determination to include the scale in their research. In a perfect world, researchers will have their pick of scales that have all been developed adequately with differing construct definitions that can provide a more fine-grained fit to their research goals. However, we recognize that this is not currently the state of the field and therefore suggest that the reader considers all aspects of the scale, including the development process, when making their choice.

In conclusion, although good definitions are the ideal, we suggest that as long as a definition is provided clearly in the text of the scale development paper, the scale can receive a point for this guideline item. Taking the time to evaluate the construct definition provides the reader with adequate information to proceed to the next step which is item evaluation.

Take home: The reader should look for a clear and precise definition of the construct. Scale developers can develop construct definitions by using a theory or data-driven approach depending on whether an agreed-upon theoretical framework of the construct exists or does not, respectively.

Question 2: Is the item generation process discussed (e.g., via a literature review, the Delphi method, crowdsourcing)? Best 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 2–3 times larger than the desired end total) that captures the construct of interest [66, 104]. 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, and crowdsourcing.

Scale developers often use a literature review to generate scale 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.

Scale developers choosing to generate items using this approach will likely already have some idea or theory about the construct in advance of the scale development process. Ideally this theory will be outlined in the scale development paper to provide clear motivation for both the stated definition and the specific items included in the initial version of the scale. As previously mentioned, a good example of a thorough literature review during the process of scale development is the one reported in Malle and Ullman [80], but there are many other examples in HRI [6, 61, 68, 108].

Another method for item generation is the Delphi method [76]. 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) [71]. This method is often implemented when the scale developer is not a subject matter expert in the construct they aim 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 final item generation method we highlight is what we refer to as the "crowdsourcing" method. Crowdsourcing 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 crowdsourcing include asking participants to sort potential items which are provided by the researchers into categories. The categories represent theoretically hypothesized factors underlying the construct of interest which have been previously determined by the researchers. An example of this approach can be found in Hoffmann et al. [61].

The crowdsourcing method may also include conducting structured interviews where participants view stimuli (e.g., a video of a robot moving through space) and are asked questions that pertain to specific components of the construct of interest. For example, in their development of a trust measure, Charalambous et al. [26] asked participants to observe a robot executing a task and then complete an interview where they were asked to discuss their thoughts about the interaction. This interview was conducted in an effort to identify relevant themes related to trust. The researchers then developed items from those themes. Alternatively, crowdsourcing can also be used to identify problematic items, such as those that may not be easily interpretable by the target population (see Section 4 for more details).

Each item development method has its respective advantages and disadvantages. If the scale developers consider themselves experts, then developing items via a literature review can be the most efficient way forward. However, experts can impose bias, even implicitly, in the types of items they choose to include in the scale. This has the potential to negatively impact the validity of the scale, as it can mean that the construct is not adequately captured by the items. If the scale developers do not consider themselves to be subject matter experts or they wish to minimize potential bias in the process, it can be very useful to outsource expertise via the Delphi method or obtain lay understanding via crowdsourcing. While these methods can potentially increase the validity of the scale by providing items for review that may not have previously been considered, it can also be difficult to find willing experts. Additionally, the iterative process, in both cases, can be time-consuming and costly. Importantly, these approaches are not mutually exclusive (i.e., any combination of methods can be used to generate items) and in some cases, combining methods can offer significant advantages.

Take home: The reader should look for any information at all (e.g., description of a procedure, pilot study reports, preliminary analyses) about how the items were generated (e.g., via literature review, Delphi method, or crowdsourcing).

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 that the scale contains some ambiguous items or that the items do not actually measure the construct as defined. For example, in the development of their trust scale, Yagoda and Gillan [124] 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). 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.

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 attitudes or attributes that the proposed construct does not encompass. For example, Malle and Ullman's [80] 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 [36]. There are many ways to formally check whether a measure has construct validity (see Section 2.3 for more details); however, this initial 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 but do not actually represent an underlying dimension that contributes to the variability in responses along the construct. In other words, initial versions of scales often include items that do not exactly fit with the defined construct. Typically, during the scale development process, specifically the item removal stage (see question 9 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. We suggest the reader personally answer each item themselves. This process can reveal critical information that may deem the scale to be not useful for the researcher's aim.[4] If the items are not easy to understand, the lack of clarity 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 [38]. 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 [34, 38].

[4]We thank an anonymous reviewer for this excellent suggestion.

Take home: The reader should 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

There are two main approaches to designing and validating a new scale: **Classical Test Theory (CTT)** and **Item Response Theory (IRT)** [37, 83, 97]. The assumption in CTT is that the participants' responses or overall score on a measure are a linear combination of their true ability with random error. The goal in CTT is to get as close to the true score as possible by minimizing the random error, or 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, or whether items and individual response patterns fit within the hypothesized model. One of the more common IRT models is the Rasch model. The Rasch model prioritizes invariance in measurement [120], or the ability of a test to measure the same construct consistently across groups or time points. The Rasch model 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 [4, 120].

While scale development methods vary, some steps are common to all. We outline these 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 [30, 78]. 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 how many factors or dimensions[5] were found, 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 CTT and IRT have different approaches for item removal, there should be some description of the criteria used and how many items were removed before the final scale is reported. Lastly, the reader should verify that 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 to determine whether the items capture the construct (i.e., guideline question 3), but also whether the sample size is large enough to determine a factor structure (i.e., guideline question 5). Additionally, the reader can determine if the factor loadings for all the items meet the relevant stated criteria (i.e., guideline question 8) and whether the scale developers removed items appropriately (i.e., guideline question 9). These are all key components of the scale development process and are related to two important tenets of science: reproducibility and replicability. Sharing the initial set of items allows others to reproduce the development process and results. This, in turn, supports the replicability of both the experimental findings and the scale's reported psychometric properties. Without the initial version of the items, it will be very difficult to determine whether the scale has been accurately developed and validated.

Take home: The reader should 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.

[5]The terms factor and dimension are used interchangeably and mean the same thing: a psychological variable that represents a component of the construct that is captured by the items within a scale.

Question 5: Does the test sample size meet the 10:1 minimum criteria? The issue of minimum adequate sample size for scale development has been extensively debated in the psychometric literature [13, 38, 40, 69, 84]. Sample sizes should be large enough to ensure that the observed item to factor relationship is stable and reliable [35, 38] and the stability of this relationship is dependent on the stability of the correlations between the items [77]. These correlations can vary across samples of all sizes, but they vary more in smaller samples than in larger ones [40]. 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 [3, 32, 78]. 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 when using different stimuli) [78].

While there is no universally agreed-upon rule for sample size within the psychometric literature, several guidelines have been suggested. For example, one class of rules suggests fixed thresholds for sample size (e.g., 300 [30, 111], 300–450 [50], and even 1,000+ [32]). Another class of rules suggests applying a ratio of items to participants (e.g., from 5:1 [45, 56, 110] up to 15:1 or 20:1 [35, 51]). Guadagnoli and Velicer [50] conducted a series of simulations which suggested that sample size should be determined based on the strength of the item to factor relationship. They suggested that strong factors can be interpreted confidently even with smaller samples, but weaker, less clearly defined factors require larger samples to ensure reliable and replicable results. Relatedly, MacCallum et al. [78] suggested that sample size should be dependent on the observed communalities (a measure of the common variance captured by items in the scale, see guideline question 8 for more details). The concern with determining sample size based on the strength of the item to factor relationship is that it is almost impossible to know *a priori* what the strength of the relationship is.

Therefore, to reconcile these recommendations, we recommend the original 10:1 rule proposed by Nunnally [92] and more recently by others [13, 87, 88]. This rule states that scale development samples should have at least 10 participants for each scale item in the initial version of the scale. However, we recognize that larger sample sizes are always better [51].

While the 10:1 rule may require quite large sample sizes in some cases, it is appropriate for developing measures in relatively more recent fields, like HRI. Many of the constructs and theories that HRI researchers are interested in are still being developed. Therefore, adopting a more conservative approach to sample size helps ensure the reliability and validity of new measures. Importantly, we recommend applying this criterion to the initial sample size where the scale is tested, since this is where the item reduction process begins.

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 with 60 items would require a sample size of at least 600 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 an opportunity for improvement for future studies in HRI. Importantly, however, if a scale does not meet this criterion, it does not necessarily mean that the scale should be discarded. It is up to the reader to determine whether the reported sample size is acceptable for their purposes and if they are comfortable with the conclusions drawn from the development process as a result.[6]

Take home: The reader should ensure that sample sizes for scale development studies follow the 10:1 (people to initial number of items) rule, though more participants are better.

[6]It is also possible to use additional statistical tests to assess the validity of a scale given the sample size *a posteriori* [121]. Of course, these results will not be accurate if the initial sample size was already too low.

Question 6: Did the scale developers perform an Exploratory Factor Analysis (EFA), Principal Components Analysis (PCA), Rasch analysis, or similar test to determine the item to factor relationship? Determining the different factors that compose a construct, as well as how those factors relate to each other is an important tenet within scientific research [36]. 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 items and the unobservable factors underlying the construct of interest. Here, factors are latent variables that are thought to explain the correlations between scale items within the construct being measured [97]. For example, Spatola et al. [108] reported four factors—sociability, agency, animacy, and disturbance—that underlie the general construct of the perception of robots in HRI. These factors suggest that robot perception is not captured by a single dimension, but by multiple, separable latent variables. Combining these variables together provides a more comprehensive measure of the overarching construct of the perception of robots.

There are many ways to investigate the relationship between items, factors, and the construct. This investigation can be completed using methods such as EFA, PCA, or the Rasch model. Most current scales are developed using one of these methods. It is important to note that PCA and factor analysis approaches (such as EFA) differ in the way they treat variance from the items, sometimes leading to different results [31, 40, 109]. 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 provide a comprehensive review of each of these methods, the interested reader can learn more about factor analysis and other methods of investigating the relationship between items and dimensions in the following resources: [43, 66, 93, 96, 97, 120] and see [12, 35, 41, 42, 46] for descriptions of the similarities and differences between factor analysis techniques and PCA. What the reader should ensure is that at the very least the study should include some description of how the scale developers determined the item to factor relationship. We refer to this step in the process as the scale development method and it should include mention conducting an EFA, PCA, or Rasch analysis.

Take home: The reader should determine whether the scale developers reported 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 multidimensional (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 [112] is a unidimensional scale. A straightforward example of a unidimensional construct is height, where an object's size can be represented by a single numerical value on a linear scale. Multidimensional scales like trust [26, 80, 113], negative attitudes toward robots [91], or perceived morality [6] are common in HRI, since historically the constructs of interest in this field have been complex and multifaceted. Multidimensional scales measure a construct along different dimensions which are then typically combined for an overall measure of the entire construct. For example, the construct of perceived morality of robots as defined by Banks [6] can be separated into two dimensions: morality and dependency. These are the two distinct, but related components of the more general construct of perceived morality. Each dimension is measured separately and then later combined to give a measure of the construct. The goal of the factor extraction process is to determine the number of factors that are necessary to describe and interpret the data.

There are a number of different ways to determine the number of factors, and each way is dependent on the scale development method used. The general process involves identifying patterns in the responses to scale items that appear to group together. This is because items that are answered

similarly, tend to reflect the same underlying (sub)factor. Then, the task is to determine how many meaningful groups appear in the data. This involves considering how the items group together as well as how much variance is explained by the groups individually and as a whole. There are many methods for identifying the number of factors (e.g., scree plots [25], parallel analysis of random data [62], and statistical tests such as chi-square test of residuals [7], very simple structure [98], or minimum average partial [115]).[7] The technical details of each of these tests is beyond the scope of the article. The reader only needs to identify whether a method to determine the number of factors was reported by the scale developers to apply this guideline criteria. If the scale developers used Rasch, a method for verifying unidimensionality (e.g., a PCA of the standardized residuals) should be discussed in the scale development paper.

Take home: The reader should verify that the scale developers reported how they determined or verified the number of factors that exist within the construct.

Question 8: Did the scale developers provide factor loadings (EFA or PCA) 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 EFA, this includes reporting factor loadings for each item. Factor loadings represent how well each item correlates with all the other items in that dimension (i.e., how well the items group together within a factor), or how much variance or covariance each latent factor is capable of explaining [40]. Higher values are better; a typical minimum factor loading is 0.6 [92].

Once the factor loadings are obtained, the factors are typically rotated so that the simplest underlying structure can be revealed. Each factor can be imagined as an axis in a coordinate system (e.g., a two-factor solution forms a two-dimensional Cartesian plane), and each item is a point in that space. The factor loading represents how closely each item aligns with a factor, or its distance from the axis [40]. In the initial, unrotated solution, the factors are extracted in order of variance explained [51], which can result in complex or even uninterpretable solutions [88]. Rotation repositions the axes relative to the items without changing the structure of the items, to create a more interpretable solution [51]. Rotation can either be orthogonal (assuming factors are uncorrelated and keeping the relationship between the axes constant) or oblique (assuming factors are correlated and allowing axes to shift closer together if necessary) [40, 51, 88].

Communality values are also sometimes reported when using factor analysis techniques. Communality refers to how much a variable shares variance with other variables. Higher communalities are better [40]. Rasch analysis uses outfit and infit measures [15, 123]. Infit is a goodness of fit statistic [122] and is the weighted average of the squared standardized residuals, where each residual is weighted by its variance [102]. Outfit, the most common method of evaluating a Rasch model, is an unweighted fit statistic, and is a measure of how well the data fit the model [102]. Outfit is sensitive to large departures from model expectations [102]. Generally, Rasch items show poor fitting items when an outfit is higher than 1.5 [75]. 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.

Additionally, we recommend that the reader not just consider the factor loadings for each item, but also evaluate how the items fit to the factor as a whole. More specifically, after reviewing factor loadings for each item, it can be beneficial to take a broader perspective to assess whether the items represent the definition or label of the factor as outlined in the paper.

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

[7]See [128] for a comparison of these methods. Also note that 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 [15, 47].

development process used an EFA or PCA), communalities, or in the form of infit/outfit values (if used 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 will not be able to capture the full scope of the construct.[8] 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 an EFA then the items might be removed based on factor loading values < 0.3 [13] or high cross-loading values (e.g., values > 0.4 of one item across two or more factors) [56]. 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 [47] or outfit < 1.3 [16]. The exact criteria used will likely vary across publications, though it is important to determine whether the criteria were applied consistently within one publication.

Items can also be removed if they are redundant with other items [38]. As mentioned in question 3, when an item is redundant, that means that there is more than one item that captures the factor to a similar level. This is distinct from removing items due to lack of fit with the construct. In that 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 case of redundant items, there is also an increased risk of additional noise but from an entirely different source, such as boredom or fatigue [59, 94]. If multiple items capture a specific aspect of the construct, it is not necessary to include all of them in the final version of the scale and therefore, scale developers may remove those redundant items. 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 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: The reader should determine whether items were removed from the final version of the scale. If items were removed, the reason (e.g., lack of fit or redundancy) should be explicitly mentioned; quantitative criteria should also be reported when possible.

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, there is an increased chance of future studies incorporating a scale that includes items that do not adequately measure to the construct. 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 in the scale development paper 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.

[8] In the unlikely event that items were not removed from the initial to final versions of a scale, the reader does not need to evaluate this guideline item.

Take home: The reader should look for the final version of the scale 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., additional EFA, **Confirmatory Factor Analysis (CFA)**, **Differential Item Functioning (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 CFA to test the factor structure. When using CFA, the latent structure uncovered during the EFA is used as a hypothesized model on a new set of data [125]. To conduct a CFA, the researcher uses the results of the initial factor analysis as a set of model parameters for a CFA. They then examine how well the CFA fits the data; most researchers will report a series of fit statistics, including RMSEA, **Tucker Lewis Index (TLI)**, CFI, and **Standardized Root Mean Square Residual (SRMR)** [19, 20, 54], though others can also be used. Each fit statistic has a heuristic value that the CFA should be under (or over). For example, Hu and Bentler [63] 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, scale developers using the Rasch method will typically focus on measurement invariance using DIF [16, 120]. DIF examines two different groups of respondents (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, such as an item that does not work as well as expected, suggesting that the item should be removed, replaced, or corrected in some way. In cases like this, the reader may find that the scale developers conducted a second CFA on a different group of participants. This second CFA examines the factor structure of the updated scale, and the same fit criteria can be applied to the results. Importantly, these tests (e.g., CFA/DIF) while valuable, represent only one step in the validation process (see guideline question 13 for more details on proper scale validation methods).

Take home: The reader should check to see if there is a test for factor structure. A CFA on a new sample or a DIF (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 measure produces similar results under similar conditions 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 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 [100]. This is because a scale can reliably measure an unintended construct (i.e., making it a reliable but not a valid measure of a construct) but it cannot be valid unless it is also reliable. In other words, reliability is a necessary, but not sufficient, prerequisite for validity.

In the context of scale development, an important component of reliability is the internal consistency of the scale. Internal consistency refers to the degree of interrelatedness among the items [33, 48]. 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 [34, 99]. ω_t is a measure of the amount of variance attributable to a general factor (the primary latent

variable) and specific factors 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 multidimensional scales, while ω_h should only be used for multidimensional scales [27].

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 the degree to which the items in a scale measure the same construct. Therefore, a high α value means that the relationships between the items account for most of the overall variability. α has been critiqued previously [28, 34, 52, 106, 127] because of its dependence on the total number of items or the assumption of tau-equivalence (i.e., that all items have the same quantitative relationship or factor loadings with the underlying construct).

Many scale developers use $\alpha \geq 0.70$ as a traditional heuristic. This often-cited standard threshold for reliability comes from Nunnally [92]. Interestingly, a closer inspection of the original text shows that this is in fact a misrepresentation. Nunnally [92] 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 research settings. This is particularly relevant in HRI researcher when results inform decisions that impact society, such as deploying robots 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. Additionally, Cronbach's α is only a valid statistic for unidimensional scales [33]. Therefore, in the case of multidimensional scales, α values should be reported for each subscale individually. For both α and ω, higher values are preferred.

Importantly, it is still crucial that the scale developers 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. Some scale developers remove an item from the scale to increase their alpha. We do not recommend this process for a number of reasons. It can lead to falsely inflated alpha values due to increased homeogeneity between the remaining items, lead to a reduction in the ability of the items to capture the construct, as well as lead to the removal of potentially meaningful measurement error that was captured by the discarded item.

Lastly, during the scale development phase, data collection methods may vary across studies. For example, some studies are conducted entirely online (e.g., [112]) while others are in-person (e.g., [68]). To our knowledge, there is no theory about the relationship between data collection methods and reliability, particularly in HRI contexts (though see [89] for a review of the impact of interaction types on scale responses in HRI contexts). The current assumption is that the results should be reliable across these different contexts and environments.

Take home: The reader should look for some test of the scale's reliability in the paper. This can be completed using metrics such as McDonald's ω_t or ω_h in addition to Cronbach's coefficient α. A reasonable minimum threshold for reliability regardless of the measure is ≥ 0.80.

Question 13: Was a test of validity (e.g., predictive, concurrent, convergent, discriminant) reported? Establishing validity is a vital step in 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 [36, 85, 96]. Importantly, and as previously mentioned, a scale cannot be valid unless it is also reliable. The concept of validity can be split into subcomponents such as criterion and construct validity [13, 36]. Criterion validity refers to the degree to which the current scale's scores relate to the same construct measured by another scale or in another context that is of interest to the scale developer [38, 96]. 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 [2]. Concurrent validity measures the degree to which the performance on the current scale relates to performance on a criterion (gold standard) measurement [36]. 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 [13]; this is especially true in HRI research.

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 [13, 17]. Construct validity can be measured in many ways [22, 29] 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 [22]. Discriminant validity refers to the extent to which the scale differs from other unrelated constructs [96]. Discriminant validity is measured by analyzing correlations between the measure of interest and other measures that do not measure the same domain or concept [13], 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. Additionally, according to the **Multitrait Multimethod (MTMM)** method [22], it is recommended to conduct at least two types of validity, for example, convergent and discriminant validity, to ensure that the construct is being adequately captured and is distinct from other, unrelated constructs (see [5] for more details on the MTMM method). It is important to note that the types of validity assessments we have surveyed here do not represent an exhaustive list of the possibilities. The interested reader should see [1] for a more comprehensive resource.

Lastly, it is often the case that, due to practical constraints (e.g., limited time and resources) 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). Additionally, validation analyses (e.g., computing a CFA or correlations for discriminant or convergent validity) can easily be integrated into ongoing research projects and these results can be reported alongside standard reliability estimates. Therefore, we strongly encourage readers to validate (e.g., by conducting a CFA) any scales they use, especially if the scale has not previously been validated. At minimum, they should explicitly acknowledge the lack of validation as a limitation in publications or presentations.

Take home: The reader should look for assessments of validity of the scale. Typically, this includes a comparison of the scale of interest to others in the field and see if there are any relationships that exist.

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. The scale developer can do this 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 guideline recommendation is that the sample size is at least 10 participants for every item in the initial version of the scale that is tested (e.g., 120 participants for a scale that contains 12 items). The reader should next identify which method was used to determine the underlying factor structure of the construct. There are many different and accepted methods for this process and 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 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 measures of reliability. Ideally, the scale developers will have computed ω_t or ω_h for the scale, depending on the dimensionality, in addition to Cronbach's coefficient α. Lastly, the reader should look for a test of validity.

Throughout the critical analysis of a scale, the reader should look for consistency in the reporting of the scale development process. More specifically, for a well-developed scale, thorough and consistent documentation should be provided for the entire process, from item generation through validation. While reviewing the scale, the reader should not be left with unanswered questions about specific aspects of the process, such as how or why items were removed or which items were included in the initial EFA. If inconsistencies exist (as they sometimes do), it should be clear exactly where and why they occurred. This transparency is crucial to ensure that the scales used in HRI are both reliable and valid measures. If the scale fails to meet these standards, its validity and overall utility should be critically reevaluated.

In addition, while pre-registration is more commonly associated with experimental studies, it can also play a valuable role in scale development. Pre-registering the scale development and validation plan, including sample size estimates, can further support transparency and strengthen the replicability of the process. Additionally, in cases where there appears to be gaps in the documentation, the reader may consider directly contacting the scale developers for further clarifications. This direct

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?	×	✓

communication can allow the reader to more efficiently make informed decisions about scale selection and interpretation rather than making assumptions. These conversations may also provide additional confidence in the process as the reader may learn more about how the scale was developed.

Armed with all of this information, the reader should now feel more confident in critically analyzing existing scales and their corresponding validation reports (see Appendix A for a glossary of terms and Appendix B and Table B1 for a printer-friendly version of the guideline). 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 article to evaluate two scales that are frequently used in HRI—the Godspeed Questionnaire [10] and the RoSAS [23]. 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 RoSAS was developed to measure the social perception of robots [23]. 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 [73, 86, 95, 117] 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

Table 2. Applying the Guideline to Two Anthropomorphism Scales

Stage	Question	Bartneck et al. [10]	Powers and Kiesler [95]
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?	×	×

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 [10]) 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 [8, 9] 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 [95] 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 [95] nor the Godspeed anthropomorphism scale [10] were adequately developed and validated.

Although the customization approaches reported in Bartneck et al. [10] to develop the Godspeed scale are not ideal, that 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 [10] 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 [23, 60, 65]). 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 [60] for a validated scale measuring some of the constructs of interest in Godspeed).

When the guideline is applied to the RoSAS we see that it meets almost all of the guideline criteria. The article 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 did not report those items, making it difficult to determine whether they adequately captured 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 article. Additionally, the sample size for the pilot studies was not adequate per the 10:1 guideline criterion (N = 210). 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 EFAs were the primary way by which both of these analyses were conducted. Results from EFA, reliability analysis, as well as the validation study (study 4) were included in the article 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 does not necessarily correlate with quality. We aim for these guidelines to empower readers to select the most suitable scale for their research question, rather than defaulting to the most commonly employed one. If the reader finds themselves needing a scale to measure constructs that are specific to Godspeed and not captured by RoSAS (e.g., animacy or perceived intelligence), we recommend seeking out newer alternatives that have undergone thorough validation. 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 researchers do when they need to measure a latent construct? The best and strongest idea is to find an existing scale and use the guideline to determine whether it has been psychometrically validated. If this process 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, 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*. The term "custom" here refers to any customization of a scale, including adaptations or modifications from existing scales. A custom scale is defined here as 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

[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 [23]. It is completely acceptable to include only a subscale in a study.

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] 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 [34]. Removing items is a nuanced process that requires expertise. This motivates us to caution against the removal of items unless further testing and validation is completed (e.g., EFA or CFA is conducted on the shortened scale). Those with the proper expertise should feel free to customize a scale and report the changes as appropriate.

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, such as adjusting tense, gender, or changing the word "automation" to "robot" [64], are usually[11] considered acceptable, whereas large changes to wording or phrasing are not. In cases where an edit is made to an existing scale, we recommend at the very least computing and reporting some measure of reliability as well as validity (e.g., convergent), where possible.

Changing the response scale (e.g., range of a Likert scale) of an existing scale is not recommended.[12] 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 (using established translation guidelines such as [53]), collect data on the translated version, and conduct a CFA using the factor structure from the originally published paper. The reader should then report the translated scale and the CFA fit indices (e.g., RMSEA, TLI, CFI) in a publication, even if the model fits do not meet the minimum criteria.

However, practical constraints (e.g., lack of resources to collect additional data) may prevent a researcher from conducting this type of analysis. In that case, the researcher might simply translate the items into 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). It is critical to include the translated list of items to ensure the replicability of the translated version of scale and increase comparability across

[10]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.

[11]We use "usually" here because, to our knowledge, there is no empirical evidence showing that the scale's robustness is affected by these minor changes. Researchers making small changes to scales are encouraged to investigate and report on the scale's reliability and validity.

[12]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 [105]. Therefore, for the purposes of the target audience we caution against the modification of response scales.

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 [114]. Therefore, we recommend researchers proceed with caution and be as transparent as possible when translating and reporting data from translated scales. (For the interested reader, see [11, 21, 39, 58, 101, 107] for more details on this process and some examples in HRI contexts.)

Relatedly, researchers may encounter situations in which they have found a "perfect" scale but a portion of their sample does not speak the language in which the scale was originally developed. In such cases, these participants may not recognize or fully understand some of the items in the scale. It is important to confirm, before testing, that participants understand all items. In online settings, this verification is typically not possible, therefore researchers should be aware of and willing to accept this risk. In clinical settings, the use of a translator can help to ensure that the intended meaning of the item(s) is preserved. However, if a scale includes items that can not be directly translated, it may not be suitable for use with that population. As previously emphasized, it is essential to consider the target population when selecting a scale to minimize these issues.[13]

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 [105] that the researcher believes best capture the latent variable of interest and then describe the modifications or item generation process. Additionally, reliability (i.e., Cronbach's alpha and McDonald's Omega) should be reported for the custom scale in the current sample, not just the reliability of the original scale. 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. Additionally, the researcher can look back in the development process to check for indicators of items that might be weaker (e.g., low factor loadings compared to the rest). When using a shortened scale, the researcher should conduct and report validation results on the new version, including an EFA/CFA and reliability. This information is relatively easy to obtain via common statistical programming software (e.g., R, SPSS) and is critical to include even if the validation of the shortened scale is not the main aim of the research study or the results of the validation do not exactly pan out. Additionally, this type of analysis may directly benefit the researchers if it validates the shortened version of the scale as a measure of the construct. The goal is to use a valid measure of a specified construct, and conducting these analyses is a crucial step in confirming its validity. Validation is an important component for scale development, regardless of whether the scale has been developed from the ground up or has been customized in some way. 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 article. The researcher should also suggest (or conduct) a follow-up study which includes a longer, validated scale to which the single item can be compared.

[13]While this guideline is not geared towards scale developers, we would like to note that we do not recommend complete avoidance of low-frequency words when developing a new scale, as doing so may dilute the construct being measured. Instead, we suggest, as a partial solution, that the scale developer consider incorporating crowdsourcing with the target population during pilot testing. This can allow for direct feedback on how items are interpreted and can help identify problematic items before formal testing.

5 Conclusions

Our aim for this tutorial was to provide a straightforward guideline for those interested in thinking critically about scales and the scale development process. Minimally, we hope that after reading through this article, 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. Alternatively, creating a custom scale might be the appropriate course of action in many cases (e.g., due to time constraints 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 (i.e., conduct an EFA/CFA and reliability measure and report the results). This is to ensure that the new scale is a valid measure of the construct 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 (or the ones that do exist have not been adequately validated), it is likely that you will need to start from the ground up and develop an entirely new measure. As previously mentioned, 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 [13, 16, 34, 70, 75, 97, 105, 120, 123].

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 article can contribute to that ideal and serve as a useful reference for any researcher interested in incorporating scales into their projects, regardless of experience level or field of research.

Acknowledgments

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] American Educational Research Association. 2014. American psychological association, and national council on measurement in education. In *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.

[2] Anne Anastasi. 1985. Psychological testing: Basic concepts and common misconceptions. In *Annual Meeting of the American Psychological Association*. American Psychological Association.

[3] J. O. Archer and R. I. Jennrich. 1973. Standard errors for orthogonally rotated factor loadings. *Psychometrika* 38 (1973), 581–592. DOI: https://doi.org/10.1007/BF02290668

[4] 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.

[5] Richard P. Bagozzi, Youjae Yi, and Lynn W. Phillips. 1991. Assessing construct validity in organizational research. *Administrative Science Quarterly* (1991), 421–458.

[6] 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.

[7] Maurice S. Bartlett. 1950. Tests of significance in factor analysis. *British Journal of Psychology*. 1950.

[8] Christoph Bartneck, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2007. Is the uncanny valley an uncanny cliff? In *16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '07)*. IEEE, 368–373.

[9] Christoph Bartneck, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. My robotic doppelgänger-A critical look at the uncanny valley. In *18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '09)*. IEEE, 269–276.

[10] 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.

[11] Dorcas E. Beaton, Claire Bombardier, Francis Guillemin, and Marcos Bosi Ferraz. 2000. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine* 25, 24 (2000), 3186–3191.

[12] Peter M. Bentler and Yutaka Kano. 1990. On the equivalence of factors and components. *Multivariate Behavioral Research* 25, 1 (1990), 67–74.

[13] 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.

[14] Kenneth A. Bollen and Rick H. Hoyle. 2012. Latent variables in structural equation modeling. In *Handbook of Structural Equation Modeling*, Rick H. Hoyle (Ed.), The Guilford Press, 56–67.

[15] T. Bond and C. Fox. 2001. *Applying the Rasch Model*. Lawrence Erlbaum Associates.

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

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

[18] John Brooke. 1996. Sus: A "quick and dirty" usability. *Usability Evaluation in Industry* 189, 3 (1996), 189–194.

[19] Timothy A. Brown. 2015. *Confirmatory Factor Analysis for Applied Research*. Guilford Publications.

[20] Timothy A. Brown and Michael T. Moore. 2012. Confirmatory factor analysis. In *Handbook of Structural Equation Modeling*. Rick H. Hoyle (Ed.), The Guilford Press, 361–379.

[21] 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.

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

[23] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In *2017 ACM/IEEE International Conference on Human–Robot Interaction*, 254–262.

[24] James McKeen Cattell. 1948. Mental tests and measurements, 1890. In *Readings in the History of Psychology*. W. Dennis (Ed.), Appleton-Century-Crofts, 347–354.

[25] Raymond B. Cattell. 1966. The scree test for the number of factors. *Multivariate Behavioral Research* 1, 2 (1966), 245–276.

[26] 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.

[27] Eunseong Cho. 2022. Reliability and omega hierarchical in multidimensional data: A comparison of various estimators. *Psychological Methods* 30 (2022), 40–59.

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

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

[30] Lee Anna Clark and David Watson. 2016. Constructing validity: Basic issues in objective scale development. In *Methodological Issues and Strategies in Clinical Research* (4th ed.). A. E. Kazdin (Ed.), American Psychological Association, 187–203.

[31] Norman Cliff. 1987. *Analyzing Multivariate Data*. Harcourt Brace Jovanovich.

[32] Andrew L. Comrey and Howard B. Lee. 2013. *A First Course in Factor Analysis*. Psychology Press.

[33] Jose M. Cortina. 1993. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology* 78, 1 (1993), 98.

[34] 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.
[35] Anna B. Costello and Jason Osborne. 2005. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation* 10 (2005), 1–9.
[36] Lee J. Cronbach and Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin* 52, 4 (1955), 281.
[37] Rafael Jaime De Ayala. 2013. *The Theory and Practice of Item Response Theory*. Guilford Publications.
[38] Robert F. DeVellis and Carolyn T. Thorpe. 2021. *Scale Development: Theory and Applications*. Sage Publications.
[39] Jessy Fenn, Chee-Seng Tan, and Sanju George. 2020. Development, validation and translation of psychological tests. *BJPsych Advances* 26, 5 (2020), 306–315.
[40] Andy Field, Zoë Field, and Jeremy Miles. 2012. *Discovering Statistics Using R*. Sage Publications.
[41] Frank J. Floyd and Keith F. Widaman. 1995. Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment* 7, 3 (1995), 286.
[42] J. Kevin Ford, Robert C. MacCallum, and Marianne Tait. 1986. The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology* 39, 2 (1986), 291–314.
[43] R. Michael Furr. 2021. *Psychometrics: An Introduction*. SAGE publications.
[44] Michael A. Goodrich and Alan C. Schultz, et al. 2008. Human–robot interaction: A survey. *Foundations and Trends® in Human–Computer Interaction* 1, 3 (2008), 203–275.
[45] Richard L. Gorsuch. 1983. *Factor Analysis* (2nd ed.). Lawrence Erlbaum Associates.
[46] Richard L. Gorsuch. 1990. Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research* 25, 1 (1990), 33–39.
[47] Kathy E. Green and Catherine G. Frantom. 2002. Survey development and validation with the Rasch model. In *International Conference on Questionnaire Development, Evaluation, and Testing*, 14–17.
[48] Samuel B. Green, Robert W. Lissitz, and Stanlen A. Mulaik. 1977. Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement* 37, 4 (1977), 827–838.
[49] 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.
[50] Edward Guadagnoli and Wayne F. Velicer. 1988. Relation of sample size to the stability of component patterns. *Psychological Bulletin* 103, 2 (1988), 265.
[51] Joseph F. Hair Jr, Wiliam C. Black, Barry J. Babin, and Rolph E. Anderson. 2010. Multivariate data analysis. *Multivariate Data Analysis (7th ed.)*, Prentice-Hall..
[52] Gregory R. Hancock and Ralph O. Mueller. 2001. Rethinking construct reliability within latent variable systems. *Structural Equation Modeling: Present and Future* 195 (2001), 216.
[53] Janet Harkness, Beth-Ellen Pennell, and Alisú Schoua-Glusberg. 2004. Survey questionnaire translation and assessment. In *Methods for Testing and Evaluating Survey Questionnaires*. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer (Eds.), Wiley, 453–473.
[54] Donna Harrington. 2009. *Confirmatory Factor Analysis*. Oxford University Press.
[55] 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.
[56] Larry Hatcher. 1994. *A Step-by-Step Approach to Using the SAS® System for Factor Analysis and Structural Equation Modeling*. SAS Institute, Cary, NC.
[57] Andrew F. Hayes and Jacob J. Coutts. 2020. Use omega rather than Cronbach's alpha for estimating reliability. But. …. *Communication Methods and Measures* 14, 1 (2020), 1–24.
[58] 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.
[59] Timothy R. Hinkin. 1995. A review of scale development practices in the study of organizations. *Journal of Management* 21, 5 (1995), 967–988.
[60] 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.
[61] Laura Hoffmann, Nikolai Bock, and Astrid M. Rosenthal Vd Pütten. 2018. The peculiarities of robot embodiment (EMCORP-scale) development, validation and initial test of the embodiment and corporeality of artificial agents scale. In *2018 ACM/IEEE International Conference on Human–Robot Interaction*, 370–378.
[62] John L. Horn. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 2 (1965), 179–185.

[63] 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.
[64] 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.
[65] 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.
[66] Paul Kline. 2013. *Handbook of Psychological Testing*. Routledge.
[67] Rex B. Kline. 2023. *Principles and Practice of Structural Equation Modeling*. Guilford Publications.
[68] 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.
[69] Theodoros A. Kyriazos, et al. 2018. Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology* 9, 08 (2018), 2207.
[70] 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.
[71] Jon Landeta. 2006. Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change* 73, 5 (2006), 467–482.
[72] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.
[73] 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.
[74] 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.
[75] John M. Linacre, M. H. Stone, J. William, P. Fisher, and L. Tesio. 2002. Rasch measurement. *Rasch Measurement Transactions* 16 (2002), 871.
[76] Harold A. Linstone and Murray Turoff. 1975. *The Delphi Method*. Addison-Wesley Reading, MA.
[77] Urbano Lorenzo-Seva and Pere J. Ferrando. 2024. Determining sample size requirements in EFA solutions: A simple empirical proposal. *Multivariate Behavioral Research* 59, 5 (2024), 899–912.
[78] Robert C. MacCallum, Keith F. Widaman, Shaobo Zhang, and Sehee Hong. 1999. Sample size in factor analysis. *Psychological Methods* 4, 1 (1999), 84.
[79] Bertram Malle. 2019. How many dimensions of mind perception really are there? In *Annual Meeting of the Cognitive Science Society*, 2268–2274.
[80] 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.
[81] Serena Marchesi, Davide Ghiglino, Francesca Ciardo, Jairo Perez-Osorio, Ebru Baykara, and Agnieszka Wykowska. 2019. Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology* 10 (2019), 450.
[82] Kate K. Mays, James J. Cummings, and James E. Katz. 2024. The robot rights and responsibilities scale: Development and validation of a metric for understanding perceptions of robots' rights and responsibilities. *International Journal of Human–Computer Interaction* (2024), 1–18.
[83] D. Betsy McCoach, Robert K. Gable, and John P. Madura. 2013. *Instrument Development in the Affective Domain*, Vol. 10, Springer.
[84] Mumtaz Ali Memon, Hiram Ting, Jun-Hwa Cheah, Ramayah Thurasamy, Francis Chuah, and Tat Huei Cham. 2020. Sample size for survey research: Review and recommendations. *Journal of Applied Structural Equation Modeling* 4, 2 (2020), i–xx.
[85] S. Messick. 1989. Validity. In *Educational Measurement*. Robert L. Linn (Ed.), American Council on Education and National Council on Measurement in Education, Washington, DC, 12–103.
[86] 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.
[87] Fabiane F. R. Morgado, Juliana F. F. Meireles, Clara M. Neves, Ana C. S. Amaral, and Maria E. C. Ferreira. 2017. Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica* 30 (2017), 3.
[88] Barbara Hazard Munro. 2005. *Statistical Methods for Health Care Research*, Vol. 1. Lippincott Williams & Wilkins.
[89] Stanislava Naneva, Marina Sarda Gou, Thomas L. Webb, and Tony J. Prescott. 2020. A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. *International Journal of Social Robotics* 12, 6 (2020), 1179–1201.
[90] John R. Nesselroade and Peter C. M. Molenaar. 2016. Some behaviorial science measurement concerns and proposals. *Multivariate Behavioral Research* 51, 2–3 (2016), 396–412.

[91] 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 *13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN '04)*. IEEE, 35–40.
[92] Jum C. Nunnally. 1978. *Psychometric Theory*. McGraw-Hill.
[93] Steven J. Osterlind. 2006. *Modern Measurement: Theory, Principles, and Applications of Mental Appraisal*. Pearson/Merrill Prentice Hall, Upper Saddle River, NJ.
[94] Philip M. Podsakoff, Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff. 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *The Journal of Applied Psychology* 88, 5 (2003), 879.
[95] Aaron Powers and Sara Kiesler. 2006. The advisor robot: Tracing people's mental model from a robot's physical attributes. In *1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, 218–225.
[96] Tenko Raykov and George A. Marcoulides. 2011. *Introduction to Psychometric Theory*. Routledge.
[97] William Revelle. 2009. An introduction to psychometric theory with applications in R. Retrieved from http://personality-project.org/r/book/
[98] William Revelle and Thomas Rocklin. 1979. Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research* 14, 4 (1979), 403–414.
[99] William Revelle and Richard E. Zinbarg. 2009. Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika* 74 (2009), 145–154.
[100] Paula Roberts and Helena Priest. 2006. Reliability and validity in research. *Nursing Standard* 20, 44 (2006), 41–46.
[101] Alexander Robitzsch and Oliver Lüdtke. 2023. Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal* 30, 6 (2023), 859–870.
[102] Peter A. M. Ruijten, Antal Haans, Jaap Ham, and Cees J. H. Midden. 2019. Perceived human-likeness of social robots: Testing the Rasch model as a method for measuring anthropomorphism. *International Journal of Social Robotics* 11 (2019), 477–494.
[103] Mark J. Schervish. 1996. P values: What they are and what they are not. *The American Statistician* 50, 3 (1996), 203–206.
[104] John A. Schinka, Wayne F. Velicer, and Irving B. Weiner. 2013. *Handbook of Psychology: Research Methods in Psychology*, Vol. 2. John Wiley & Sons, Inc.
[105] 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.
[106] Klaas Sijtsma. 2009. On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika* 74 (2009), 107–120.
[107] 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.
[108] Nicolas Spatola, Barbara Kühnlenz, and Gordon Cheng. 2021. Perception and evaluation in human–robot interaction: The human–robot interaction evaluation scale (HRIES)—a multicomponent approach of anthropomorphism. *International Journal of Social Robotics* 13, 7 (2021), 1517–1539.
[109] James Stevens. 2002. *Applied Multivariate Statistics for the Social Sciences*, Vol. 4. Lawrence Erlbaum Associates, Mahwah, NJ.
[110] Diana D. Suhr. 2006. Exploratory or confirmatory factor analysis. In *Proceedings of SAS Users Group International Conference*, 1–17.
[111] Barbara G. Tabachnick, Linda S. Fidell, and J. B. Ullman. 2019. *Using Multivariate Statistics* (6th ed.), Pearson Education, 497–516.
[112] 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 *32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN '23)*, 936–942. DOI: https://doi.org/10.1109/RO-MAN57019.2023.10309544
[113] 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.
[114] Fons Van de Vijver and Kwok Leung. 1997. Methods and data analysis of comparative research. In *Handbook of Cross-Cultural Psychology* (2nd ed.). J. W. Berry, Y. H. Poortinga, and J. Pandey (Eds.), Allyn & Bacon, 257–300.
[115] Wayne F. Velicer. 1976. The relation between factor score estimates, image scores, and principal component scores. *Educational and Psychological Measurement* 36, 1 (1976), 149–159.
[116] Bertie Vidgen and Taha Yasseri. 2016. P-values: Misunderstood and misused. *Frontiers in Physics* 4 (2016), 6.
[117] 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.

[118] Ronald L. Wasserstein and Nicole A. Lazar. 2016. The ASA statement on p-values: Context, process, and purpose. *The American Statistician* 70 (2016), 129–133.
[119] 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.
[120] Stefanie Wind and Cheng Hua. 2021. Rasch measurement theory analysis in R: Illustrations and practical guidance for researchers and practitioners. Bookdown.org.
[121] Erika J. Wolf, Kelly M. Harrington, Shaunna L. Clark, and Mark W. Miller. 2013. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement* 73, 6 (2013), 913–934.
[122] Benjamin D. Wright and Magdalena M. C. Mok. 2004. An overview of the family of Rasch measurement models. *Introduction to Rasch Measurement* 1, 1 (2004), 1–24.
[123] Benjamin D. Wright and Mark H. Stone. 1979. Best test design. MESA Press.
[124] 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.
[125] Matthias Ziegler and Dirk Hagemann. 2015. Testing the unidimensionality of items. *European Journal of Psychological Assessment* 31 (2015), 231–237.
[126] 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.
[127] 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.
[128] William R. Zwick and Wayne F. Velicer. 1986. Comparison of five rules for determining the number of components to retain. *Psychological Bulletin* 99, 3 (1986), 432.

Appendices

A Glossary

Classical Test Theory (CTT). A scale development method where the assumption 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 noise.

Communality. A measure of the common variance captured by items in the scale.

Confirmatory Factor Analysis (CFA). A test of the factor structure of a scale. Typically performed after an exploratory factor analysis (EFA). When using CFA, the latent structure uncovered during the exploratory factor analysis is used as a hypothesized model on a new set of data. Fit indices are used to determine model fit.

Construct. A construct refers to the unobserved (i.e., latent) attitude, cognition, or attribute that is the target of the study. Unobserved (or latent) in this context simply refers to a type of construct that exists in the mind of the participant and cannot easily be directly observed. The term construct can be used interchangeably with other terms such as domain and latent variable.

Construct Validity. Construct validity 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.

Convergent Validity. Convergent validity refers to how well the new scale correlates with other variables that are designed to measure similar constructs.

Cronbach's Coefficient Alpha (α). A reliability measure. α is a measure of the degree to which items in the scale measure the same construct. The value is based on the average interitem covariance.

Dimension. A psychological variable that represents a component of the construct that is captured by the items within a scale. This term is also used interchangeably with factor.
Discriminant Validity. Discriminant validity refers to the extent to which the scale differs from other unrelated constructs. Discriminant validity is measured by analyzing correlations between the

measure of interest and other measures that do not measure the same domain or concept, where weaker correlations are expected.

Exploratory Factor Analysis (EFA). A method for determining the structure of latent variables and their relationships within a construct.

Factor. A psychological variable that represents a component of the construct that is captured by the items within a scale. This term is also used interchangeably with dimension.

Factor Loading. A value that represents how well each item correlates with all the other items in that dimension (i.e., how well items group together within a factor), or how much variance or covariance each latent factor is capable of explaining.

Infit/Outfit. Methods for assessing item fit in Rasch analysis. Infit is a goodness of fit statistic and is the weighted average of the squared standard residuals where each residual is weighted by its variance. Outfit is an unweighted fit statistic and is a measure of how well the data fit the model.

Item. An item refers to the direct questions, directives, or statements that make up a scale. Each item within a scale is intended to capture the construct (i.e., attitude or behavior) either in part or in full.

Item Response Theory (IRT). Item response theory uses an item-level approach to determining item and person fit within the scale.

Latent Variable. The unobservable behavior, attitude, or attribute that is being measured. This term is often used interchangeably with construct and domain.

McDonald's Omega (ω). A measure of how reliably a set of items measures a single factor or construct. Omega total (ω_t) is a measure of the amount of variance attributable to a general factor (the primary latent variable) and specific factors (items). Omega hierarchical (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 multidimensional scales, while ω_h should only be used for multidimensional scales.

Rasch. One of the more common IRT models is the Rasch model. The Rasch model prioritizes invariance in measurement 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.

Reliability. Reliability refers to the principle that a measurement produces similar results under similar conditions.

Rotation. A method of rotating factor axes such that variables (items) load maximally onto factors. Rotation repositions the axes relative to the items without changing the structure of the items. Rotation can either be orthogonal (assuming factors are uncorrelated) or oblique (assuming factors are correlated).

Scale. The term "scale" refers to any instrument that measures a behavior, attitude, or other latent construct that is not directly observable.

Subscale. 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.

Received 26 January 2024; revised 17 July 2025; accepted 9 September 2025

B Printer-Friendly Version of the Guideline

Table B1. 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 crowdsourcing)?	□
	(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/PCA) 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., additional 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)?	□