



Chapter 3: Understanding Test Quality-Concepts of Reliability and Validity

Test *reliability* and *validity* are two technical properties of a test that indicate the quality and usefulness of the test. These are the two most important features of a test. You should examine these features when evaluating the suitability of the test for your use. This chapter provides a simplified explanation of these two complex ideas. These explanations will help you to understand reliability and validity information reported in test manuals and reviews and use that information to evaluate the suitability of a test for your use.

Chapter Highlights

1. What makes a good test?
2. Test reliability
3. Interpretation of reliability information from test manuals and reviews
4. Types of reliability estimates
5. Standard error of measurement
6. Test validity
7. Methods for conducting validation studies
8. Using validity evidence from outside studies
9. How to interpret validity information from test manuals and independent reviews.

Principles of Assessment Discussed Use only *reliable* assessment instruments and procedures. Use only assessment procedures and instruments that have been demonstrated to be valid for the specific purpose for which they are being used. Use assessment tools that are appropriate for the target population.

What makes a good test?

An employment test is considered "good" if the following can be said about it:

- The test measures what it claims to measure consistently or reliably. This means that if a person were to take the test again, the person would get a *similar* test score.
- The test measures what it claims to measure. For example, a test of mental ability does in fact measure mental ability, and not some other characteristic.
- The test is job-relevant. In other words, the test measures one or more characteristics that are important to the job.
- By using the test, more effective employment decisions can be made about individuals. For example, an arithmetic test may help you to select qualified workers for a job that requires knowledge of arithmetic operations.

The degree to which a test has these qualities is indicated by two technical properties: *reliability* and *validity*.

Test reliability

Reliability refers to how dependably or consistently a test measures a characteristic. If a person takes the test again, will he or she get a similar test score, or a much different score? A test that yields similar

scores for a person who repeats the test is said to measure a characteristic reliably.

How do we account for an individual who does not get exactly the same test score every time he or she takes the test? Some possible reasons are the following:

- **Test taker's temporary psychological or physical state.** Test performance can be influenced by a person's psychological or physical state at the time of testing. For example, differing levels of anxiety, fatigue, or motivation may affect the applicant's test results.
- **Environmental factors.** Differences in the testing environment, such as room temperature, lighting, noise, or even the test administrator, can influence an individual's test performance.
- **Test form.** Many tests have more than one version or form. Items differ on each form, but each form is supposed to measure the same thing. Different forms of a test are known as **parallel forms or alternate forms**. These forms are designed to have similar measurement characteristics, but they contain different items. Because the forms are not exactly the same, a test taker might do better on one form than on another.
- **Multiple raters.** In certain tests, scoring is determined by a rater's judgments of the test taker's performance or responses. Differences in training, experience, and frame of reference among raters can produce different test scores for the test taker.

These factors are sources of chance or random measurement error in the assessment process. If there were no random errors of measurement, the individual would get the same test score, the individual's "true" score, each time. The degree to which test scores are unaffected by measurement errors is an indication of the **reliability of the test**.

Principle of Assessment: Use only *reliable* assessment instruments and procedures. In other words, use only assessment tools that provide dependable and consistent information.

Reliable assessment tools produce dependable, repeatable, and consistent information about people. In order to meaningfully interpret test scores and make useful employment or career-related decisions, you need reliable tools. This brings us to the next principle of assessment.

Interpretation of reliability information from test manuals and reviews

Test manuals and independent review of tests provide information on test reliability. The following discussion will help you interpret the reliability information about any test.

The reliability of a test is indicated by the **reliability coefficient**. It is denoted by the letter "r," and is expressed as a number ranging between 0 and 1.00, with $r = 0$ indicating no reliability, and $r = 1.00$ indicating perfect reliability. Do not expect to find a test with perfect reliability. Generally, you will see the reliability of a test as a decimal, for example, $r = .80$ or $r = .93$. The larger the reliability coefficient, the more repeatable or reliable the test scores. Table 1 serves as a general guideline for interpreting test reliability. However, do **not** select or reject a test solely based on the size of its reliability coefficient. To evaluate a test's reliability, you should consider the type of test, the type of reliability estimate reported, and the context in which the test will be used.

Table 1. General Guidelines for

Reliability coefficient value	Interpretation
.90 and up	excellent
.80 - .89	good
.70 - .79	adequate
below .70	may have limited applicability

Types of reliability estimates

There are several types of reliability estimates, each influenced by different sources of measurement error. Test developers have the responsibility of reporting the reliability estimates that are relevant for a particular test. Before deciding to use a test, read the test manual and any independent reviews to determine if its reliability is acceptable. The acceptable level of reliability will differ depending on the type of test and the reliability estimate used.

The discussion in Table 2 should help you develop some familiarity with the different kinds of reliability estimates reported in test manuals and reviews.

Table 2. Types of Reliability Estimates

- **Test-retest reliability** indicates the repeatability of test scores with the passage of time. This estimate also reflects the stability of the characteristic or construct being measured by the test.

Some constructs are more stable than others. For example, an individual's reading ability is more stable over a particular period of time than that individual's anxiety level. Therefore, you would expect a higher test-retest reliability coefficient on a reading test than you would on a test that measures anxiety. For constructs that are expected to vary over time, an acceptable test-retest reliability coefficient may be lower than is suggested in Table 1.

- **Alternate or parallel form reliability** indicates how consistent test scores are likely to be if a person takes two or more forms of a test.

A high parallel form reliability coefficient indicates that the different forms of the test are very similar which means that it makes virtually no difference which version of the test a person takes. On the other hand, a low parallel form reliability coefficient suggests that the different forms are probably **not** comparable; they may be measuring different things and therefore cannot be used interchangeably.

- **Inter-rater reliability** indicates how consistent test scores are likely to be if the test is scored by two or more raters.

On some tests, raters evaluate responses to questions and determine the score. Differences in judgments among raters are likely to produce variations in test scores. A high inter-rater reliability coefficient indicates that the judgment process is stable and the resulting scores are reliable.

Inter-rater reliability coefficients are typically lower than other types of reliability estimates. However, it is possible to obtain higher levels of inter-rater reliabilities if raters are appropriately trained.

- **Internal consistency reliability** indicates the extent to which items on a test measure the same thing.

A high internal consistency reliability coefficient for a test indicates that the items on the test are very similar to each other in content (homogeneous). It is important to note that the length of a test can affect internal consistency reliability. For example, a very lengthy test can spuriously inflate the reliability coefficient.

Tests that measure multiple characteristics are usually divided into distinct components. Manuals for such tests typically report a separate internal consistency reliability coefficient for each component in addition to one for the whole test.

Test manuals and reviews report several kinds of internal consistency reliability estimates. Each type of estimate is appropriate under certain circumstances. The test manual should explain why a particular estimate is reported.

Standard error of measurement

Test manuals report a statistic called the **standard error of measurement (SEM)**. It gives the margin of error that you should expect in an individual test score because of imperfect reliability of the test. The SEM represents the degree of confidence that a person's "true" score lies within a particular range of scores. For example, an SEM of "2" indicates that a test taker's "true" score probably lies within 2 points

in either direction of the score he or she receives on the test. This means that if an individual receives a 91 on the test, there is a good chance that the person's "true" score lies somewhere between 89 and 93.

The SEM is a useful measure of the accuracy of individual test scores. The smaller the SEM, the more accurate the measurements.

When evaluating the reliability coefficients of a test, it is important to review the explanations provided in the manual for the following:

- **Types of reliability used.** The manual should indicate why a certain type of reliability coefficient was reported. The manual should also discuss sources of random measurement error that are relevant for the test.
- **How reliability studies were conducted.** The manual should indicate the conditions under which the data were obtained, such as the length of time that passed between administrations of a test in a test-retest reliability study. In general, reliabilities tend to drop as the time between test administrations increases.
- **The characteristics of the sample group.** The manual should indicate the important characteristics of the group used in gathering reliability information, such as education level, occupation, etc. This will allow you to compare the characteristics of the people you want to test with the sample group. If they are sufficiently similar, then the reported reliability estimates will probably hold true for your population as well.

For more information on reliability, consult the APA Standards, the SIOP Principles, or any major textbook on psychometrics or employment testing. Appendix A lists some possible sources.

Test validity

Validity is the most important issue in selecting a test. Validity refers to *what characteristic* the test measures and *how well* the test measures that characteristic.

- Validity tells you if the characteristic being measured by a test is related to job qualifications and requirements.
- Validity gives *meaning* to the test scores. Validity evidence indicates that there is linkage between test performance and job performance. It can tell you what you may conclude or predict about someone from his or her score on the test. If a test has been demonstrated to be a valid predictor of performance on a specific job, you can conclude that persons scoring high on the test are more likely to perform well on the job than persons who score low on the test, all else being equal.
- Validity also describes the degree to which you can make specific conclusions or predictions about people based on their test scores. In other words, it indicates the usefulness of the test.

It is important to understand the differences between *reliability* and *validity*. Validity will tell you how good a test is for a particular situation; reliability will tell you how trustworthy a score on that test will be. You cannot draw valid conclusions from a test score unless you are sure that the test is reliable. Even when a test is reliable, it may not be valid. You should be careful that any test you select is both reliable and valid for your situation.

Principle of Assessment: Use only assessment procedures and instruments that have been demonstrated to be valid for the specific purpose for which they are being used.

A test's validity is established in reference to a specific purpose; the test may not be valid for different purposes. For example, the test you use to make valid predictions about someone's technical proficiency on the job may not be valid for predicting his or her leadership skills or absenteeism rate. This leads to the next principle of assessment.

Similarly, a test's validity is established in reference to specific groups. These groups are called the reference groups. The test may not be valid for different groups. For example, a test designed to predict the performance of managers in situations requiring problem solving may not allow you to make valid or meaningful predictions about the performance of clerical employees. If, for example, the kind of problem-solving ability required for the two positions is different, or the reading level of the test is not suitable for clerical applicants, the test results may be valid for managers, but not for clerical employees.

Test developers have the responsibility of describing the reference groups used to develop the test. The manual should describe the groups for whom the test is valid, and the interpretation of scores for

individuals belonging to each of these groups. You must determine if the test can be used appropriately with the particular type of people you want to test. This group of people is called your *target population* or *target group*.

Your target group and the reference group do **not** have to match on all factors; they must be sufficiently similar so that the test will yield meaningful scores for your group.

For example, a writing ability test

developed for use with college seniors may

be appropriate for measuring the writing

ability of white-collar professionals or managers, even though these groups do not have identical

characteristics. In determining the appropriateness of a test for your target groups, consider factors such as occupation, reading level, cultural differences, and language barriers.

Principle of Assessment: Use assessment tools that are appropriate for the target population.

Recall that the *Uniform Guidelines* require assessment tools to have adequate supporting evidence for the conclusions you reach with them in the event adverse impact occurs. A valid personnel tool is one that measures an important characteristic of the job you are interested in. Use of valid tools will, on average, enable you to make better employment-related decisions. Both from business-efficiency and legal viewpoints, it is essential to only use tests that are valid for your intended use.

In order to be certain an employment test is useful and valid, evidence must be collected relating the test to a job. The process of establishing the job relatedness of a test is called **validation**.

Methods for conducting validation studies

The *Uniform Guidelines* discuss the following three methods of conducting validation studies. The *Guidelines* describe conditions under which each type of validation strategy is appropriate. They do not express a preference for any one strategy to demonstrate the job-relatedness of a test.

- **Criterion-related validation** requires demonstration of a correlation or other statistical relationship between test performance and job performance. In other words, individuals who score high on the test tend to perform better on the job than those who score low on the test. If the criterion is obtained at the same time the test is given, it is called concurrent validity; if the criterion is obtained at a later time, it is called predictive validity.
- **Content-related validation** requires a demonstration that the content of the test represents important job-related behaviors. In other words, test items should be relevant to and measure directly important requirements and qualifications for the job.
- **Construct-related validation** requires a demonstration that the test measures the construct or characteristic it claims to measure, and that this characteristic is important to successful performance on the job.

The three methods of validity-criterion-related, content, and construct-should be used to provide validation support depending on the situation. These three general methods often overlap, and, depending on the situation, one or more may be appropriate. French (1990) offers situational examples of when each method of validity may be applied.

First, as an example of criterion-related validity, take the position of millwright. Employees' scores (predictors) on a test designed to measure mechanical skill could be correlated with their performance in servicing machines (criterion) in the mill. If the correlation is high, it can be said that the test has a high degree of validation support, and its use as a selection tool would be appropriate.

Second, the content validation method may be used when you want to determine if there is a relationship between behaviors measured by a test and behaviors involved in the job. For example, a typing test would be high validation support for a secretarial position, assuming much typing is required each day. If, however, the job required only minimal typing, then the same test would have little content validity. Content validity does not apply to tests measuring learning ability or general problem-solving skills (French, 1990).

Finally, the third method is construct validity. This method often pertains to tests that may measure abstract traits of an applicant. For example, construct validity may be used when a bank desires to test its applicants for "numerical aptitude." In this case, an aptitude is not an observable behavior, but a concept created to explain possible future behaviors. To demonstrate that the test possesses construct validation support, ". . . the bank would need to show (1) that the test did indeed measure the desired trait and (2) that this trait corresponded to success on the job" (French, 1990, p. 260).

Professionally developed tests should come with reports on validity evidence, including detailed explanations of how validation studies were conducted. If you develop your own tests or procedures, you will need to conduct your own validation studies. As the test user, you have the ultimate responsibility for making sure that validity evidence exists for the conclusions you reach using the tests. This applies to all tests and procedures you use, whether they have been bought off-the-shelf, developed externally, or developed in-house.

Validity evidence is especially critical for tests that have adverse impact. When a test has adverse impact, the *Uniform Guidelines* require that validity evidence for that specific employment decision be provided.

The particular job for which a test is selected should be very similar to the job for which the test was originally developed. Determining the degree of similarity will require a **job analysis**. Job analysis is a systematic process used to identify the tasks, duties, responsibilities and working conditions associated with a job and the knowledge, skills, abilities, and other characteristics required to perform that job.

Job analysis information may be gathered by direct observation of people currently in the job, interviews with experienced supervisors and job incumbents, questionnaires, personnel and equipment records, and work manuals. In order to meet the requirements of the *Uniform Guidelines*, it is advisable that the job analysis be conducted by a qualified professional, for example, an industrial and organizational psychologist or other professional well trained in job analysis techniques. Job analysis information is central in deciding what to test for and which tests to use.

Using validity evidence from outside studies

Conducting your own validation study is expensive, and, in many cases, you may not have enough employees in a relevant job category to make it feasible to conduct a study. Therefore, you may find it advantageous to use professionally developed assessment tools and procedures for which documentation on validity already exists. However, care must be taken to make sure that validity evidence obtained for an "outside" test study can be suitably "transported" to your particular situation.

The *Uniform Guidelines*, the *Standards*, and the *SIOP Principles* state that evidence of transportability is required. Consider the following when using outside tests:

- **Validity evidence.** The validation procedures used in the studies must be consistent with accepted standards.
- **Job similarity.** A job analysis should be performed to verify that your job and the original job are substantially similar in terms of ability requirements and work behavior.
- **Fairness evidence.** Reports of test fairness from outside studies must be considered for each protected group that is part of your labor market. Where this information is not available for an otherwise qualified test, an internal study of test fairness should be conducted, if feasible.
- **Other significant variables.** These include the type of performance measures and standards used, the essential work activities performed, the similarity of your target group to the reference samples, as well as all other situational factors that might affect the applicability of the outside test for your use.

To ensure that the outside test you purchase or obtain meets professional and legal standards, you should consult with testing professionals. See [Chapter 5](#) for information on locating consultants.

How to interpret validity information from test manuals and independent reviews

To determine if a particular test is valid for your intended use, consult the test manual and available independent reviews. ([Chapter 5](#) offers sources for test reviews.) The information below can help you interpret the validity evidence reported in these publications.

In evaluating validity information, it is important to determine whether the test can be used in the specific way you intended, and whether your target group is similar to the test reference group.

Test manuals and reviews should describe

- Available validation evidence supporting use of the test for specific purposes. The manual should include a thorough description of the procedures used in the validation studies and the results of

those studies.

- The possible valid uses of the test. The purposes for which the test can legitimately be used should be described, as well as the performance criteria that can validly be predicted.
- The sample group(s) on which the test was developed. For example, was the test developed on a sample of high school graduates, managers, or clerical workers? What was the racial, ethnic, age, and gender mix of the sample?
- The group(s) for which the test may be used.

The *criterion-related validity* of a test is measured by the *validity coefficient*. It is reported as a number between 0 and 1.00 that indicates the magnitude of the relationship, "r," between the test and a measure of job performance (criterion). The larger the validity coefficient, the more confidence you can have in predictions made from the test scores. However, a single test can never fully predict job performance because success on the job depends on so many varied factors. Therefore, validity coefficients, unlike reliability coefficients, rarely exceed $r = .40$.

Table 3. General Guidelines for Interpreting Validity Coefficients

As a general rule, the higher the validity coefficient the more beneficial it is to use the test. Validity coefficients of $r = .21$ to $r = .35$ are typical for a single test. Validities for selection systems that use multiple tests will probably be higher because you are using different tools to measure/predict different aspects of performance, where a single test is more likely to measure or predict fewer aspects of total performance. Table 3 serves as a general guideline for interpreting test validity for a single test. Evaluating test validity is a sophisticated task, and you might require the services of a testing expert. In addition to the magnitude of the validity coefficient, you should also consider at a minimum the following factors:

- level of adverse impact associated with your assessment tool
- selection ratio (number of applicants versus the number of openings)
- cost of a hiring error
- cost of the selection tool
- probability of hiring qualified applicant based on chance alone.

Validity coefficient value	Interpretation
above .35	very beneficial
.21 - .35	likely to be useful
.11 - .20	depends on circumstances
below .11	unlikely to be useful

Here are three scenarios illustrating why you should consider these factors, individually and in combination with one another, when evaluating validity coefficients:

Scenario One

You are in the process of hiring applicants where you have a high selection ratio and are filling positions that do not require a great deal of skill. In this situation, you might be willing to accept a selection tool that has validity considered "likely to be useful" or even "depends on circumstances" because you need to fill the positions, you do not have many applicants to choose from, and the level of skill required is not that high.

Now, let's change the situation.

Scenario Two

You are recruiting for jobs that require a high level of accuracy, and a mistake made by a worker could be dangerous and costly. With these additional factors, a slightly lower validity coefficient would probably not be acceptable to you because hiring an unqualified worker would be too much of a risk. In this case you would probably want to use a selection tool that reported validities considered to be "very beneficial" because a hiring error would be too costly to your company.

Here is another scenario that shows why you need to consider multiple factors when evaluating the validity of assessment tools.

Scenario Three

A company you are working for is considering using a very costly selection system that results in fairly high levels of adverse impact. You decide to implement the selection tool because the assessment tools you found with lower adverse impact had substantially lower validity, were just as costly, and making mistakes in hiring decisions would be too much of a risk for your company. Your company decided to implement the assessment given the difficulty in hiring for the particular positions, the "very beneficial"

validity of the assessment and your failed attempts to find alternative instruments with less adverse impact. However, your company will continue efforts to find ways of reducing the adverse impact of the system.

Again, these examples demonstrate the complexity of evaluating the validity of assessments. Multiple factors need to be considered in most situations. You might want to seek the assistance of a testing expert (for example, an industrial/organizational psychologist) to evaluate the appropriateness of particular assessments for your employment situation.

When properly applied, the use of valid and reliable assessment instruments will help you make better decisions. Additionally, by using a variety of assessment tools as part of an assessment program, you can more fully assess the skills and capabilities of people, while reducing the effects of errors associated with any one tool on your decision making.

A document by the:

[U.S. Department of Labor
Employment and Training Administration
1999](#)



Ten Steps in Scale Development and Reporting: A Guide for Researchers

Serena Carpenter 

School of Journalism, Michigan State University, East Lansing, MI, USA

ABSTRACT

Scale development involves numerous theoretical, methodological, and statistical competencies. Despite the central role that scales play in our predictions, scholars often apply measurement building procedures that are inconsistent with best practices. The defaults in statistical programs, inadequate training, and numerous evaluation points can lead to improper practices. Based on a quantitative content analysis of communication journal articles, scholars have improved very little in the communication of their scale development decisions and practices. To address these reoccurring issues, this article breaks down and recommends 10 steps to follow in the scale development process for researchers unfamiliar with the process. Furthermore, the present research makes a unique contribution by over-viewing procedures scholars should employ to develop their dimensions and corresponding items. The overarching objective is to encourage the adoption of scale development best practices that yield stronger concepts, and in the long run, a more stable foundation of knowledge.

Social scientific terms such as interactivity, source expertise, and media credibility organize our thinking about research. Theoreticians construct abstract measures based on the collective scientific community's interpretation of a latent term. The hallmark of the quantitative approach is that concepts have been grounded in systematic procedures that enable scholars to apply their measures in similar or varying settings to determine their usefulness. Usefulness of a particular concept is often determined by the concept's ability to predict phenomena and make claims of scientific knowledge, but that knowledge may be imprecise if scholars are not aware of proper scale development techniques and reporting procedures. The linking of measurement indicators to a concept is a complex process. Unfortunately, literature on measurement theory and practice to guide communication scholars is not emphasized.

The purpose of the present article is to be constructive by informing readers about the misuses of Exploratory Factor Analysis (EFA) in order to discourage such mistakes in the future. A lack of awareness regarding the appropriate procedures has prompted leading scale methodologists to argue that most measures are seriously flawed even in reputable journals (Conway & Huffcutt, 2003; Kline, 2013; McCrosky & Young, 1979). Furthermore, three separate content analyses of communication journal articles published from 1978–2009 found that statistical and methodological decisions associated with scale development were poor, which provides evidence concerning the existence of questionable measures in the field of communication (Morrison, 2009; Park, Dailey, & Lemus, 2002; Wimmer & Haynes, 1978).

Scholars learn formal measurement development expectations and our dedication to measurement quality through the observation of what scientists do. Scale development is not often taught in doctoral programs, which likely means that scholars learn by imitating procedures communicated in

research journals (Conway & Huffcutt, 2003; Ford, MacCullum, & Tait, 1986; Park et al., 2002). This article examined the practices of scholars that applied procedures to develop their scales in top communication journals ranked by Thomson Reuter representing a ten-year period from 2005–2015 to determine whether the findings from previous content analyses hold in a more recent data set. It did not evaluate the work of scholars who created scales without the use of factor analysis procedures. The present study also extended previous content analysis work through the examination of several other variables essential to the measurement building process (e.g., the scale item generation and assessment procedures, sample characteristics, appropriateness of FA, item deletion process, factor cut-off levels, etc.). Additionally, communication and media researchers' concept explication choices concerning how they built their scales decisions *prior* to the application of statistics and methods were reviewed. Based on the present ($n = 600$) and previous content analysis results that continue to demonstrate that authors do not abide by scale development best practices, this article enacts a narrative intended to support researchers by breaking down the scale development process into ten manageable steps (see Table 1).

Scale development concepts

Scales try to capture not directly observable latent concepts with a group of concrete statements. Scales are “collections of items combined into a composite score intended to reveal levels of

Table 1. 10 steps in scale development and reporting.

1. Research the intended meaning and breadth of the theoretical concept
a. Select appropriate conceptual labels
b. Select conceptual definitions
c. Identify potential dimensions and items
d. Conduct qualitative research to generate dimensions and items
i. Use feedback to refine scale
i. Expert feedback, pre-tests, cognitive interviews, or pilot tests can be employed to evaluate item wording, item validity, questionnaire design, and model structure
2. Determine sampling procedure (5:1)
3. Examine data quality
4. Verify the factorability of the data
a. Bartlett's Test of Sphericity ($\leq .05$)
b. Kaiser-Meyer-Olkin test of sampling adequacy ($\geq .60$)
c. Inspect correlation matrix ($\geq .30$)
5. Conduct Common Factor Analysis
6. Select factor extraction method
a. Principal Factors Analysis
b. Maximum Likelihood
7. Determine number of factors
a. Theoretical convergence and parsimony
b. Scree test
c. Parallel Analysis (PA)
d. Minimum Average Partial (MAP)
8. Rotate factors
a. Oblique rotation (Direct Oblimin, Promax)
9. Evaluate items based on a priori criteria
a. Theoretical convergence
b. Parsimony
c. Weak loadings ($\geq .32$)
d. Cross loadings
e. Inter-item correlations
f. At least three-item factors
g. Communalities of items ($\geq .40$)
10. Present results
a. Scale and subscale naming logic, conceptual definitions, sample size logic, methods for determining factor numbers, Bartlett's test of sphericity, Kaiser-Meyer-Olkin test of sampling adequacy results, factor extraction method, rotational method, strategies for deciding on items, eigenvalues for all factors, pattern matrix, computer program package, communalities for each variable, descriptive statistics, subscale reliabilities, and percentage of variance accounted for by each factor.

theoretical variables not readily observable by direct means (DeVellis, 2012, p. 11).” Scholars are not able to observe the direct relationship among items, but they can determine if they are sufficiently intercorrelated with one another (DeVellis, 2012). If a scientific construct is truly abstract, subscales should comprise of at least three variables to capture the true central of the concept and to ensure content validity (Viswanathan, 2010). Multiple empirical items protect against the influence of culture, biases, and item order (Morrison, 2009).

Scale dimensions (i.e., subscales, factors)

Researchers also need to investigate whether the latent variable is a unidimensional or a multidimensional measure. If it is multidimensional, the scale will need to be eventually split into subscales that represent one composite scale. In fact, 70% of measures published in a *Psychological Assessment* sample included subscales (Clark & Watson, 1995). A literature review is necessary to map the dimensional structure of the construct because researchers need to craft items that reflect their theoretical understanding of each dimension. As abstractness (and breadth) increases, one can expect the construct to be comprised of more than one dimension. For example, a literature review of *graduate student mentoring support* may reveal three separate types (or dimensions) of faculty mentoring support: psychosocial, research, and career.

Exploratory factor analysis

Following the literature review, conceptual definition proposal, and exploratory methodological work to identify dimensions and items, exploratory factor analysis (EFA) is the most often applied approach in evaluating proposed scales. Factor analysis is a group of structure analyzing procedures used to identify correlations among observable variables to aid in the data reduction of variables related to each dimension (i.e., factor) of the construct (Norris & Lecavalier, 2010). Essentially, EFA explores the data and provides guidance on factor number. In a confirmatory factor analysis (CFA), researchers specify factor number and the associated variables with each factor prior to conducting one. EFA is recommended over CFA for scale development due to the possibility that researchers are incorrect regarding their assumptions about the construct’s dimensionality and to also ensure item quality. A CFA should be conducted on a separate sample to *confirm* the structure of the proposed scale resulting from an EFA (Costello & Osborne, 2005; Ford et al., 1986; Haig, 2005; Kline, 2013; Pett, Lackey, & Sullivan, 2003; Preacher & MacCallum, 2003; Worthington & Whittaker, 2006). Scholars should never assume the rigor of published scales. All published scales should be submitted to a confirmatory factor analysis to validate the dimensional structure of a measure in order to prevent large bodies of literature being built on invalid scales (Levine, Hullett, Turner, & Lapinski, 2006). The complexity of the *scale development* and *scale validation* process can result in several missteps, but *scale development* is the informational focus of this article.

Common problems and issues

Today, factor analysis decisions are becoming an even more prominent issue as the number of communication scholars using factor analysis is increasing (Park et al., 2002; Ye & Ki, 2012). Preacher and MacCallum (2003) argued the most important decisions in scale development include deciding between common factor analysis and PCA, the number of dimensions (i.e., factors) to retain, and the rotational method (oblique vs. orthogonal). The default functions in many statistical analysis programs such as SPSS or SAS lead many researchers to incorrectly utilize techniques such as principal components analysis, Varimax rotation, and eigenvalues greater than one (Conway & Huffcutt, 2003; Park et al., 2002; Reise, Waller, & Comrey, 2000). Kaiser (1970) referred to this three-pronged approach to factor analysis as the *Little Jiffy*. The *Little Jiffy* approach may be the norm in communication research; however, it does not yield precise results (Costello & Osborne, 2005).

Content analysis results of journal articles outside and inside the field of communication have focused on the choices of researchers finding that factor analysis is one of the most misunderstood procedures in the social sciences. Exploratory factor analyses have been critiqued in the fields of developmental disabilities, organizational research, counseling psychology, and psychology. The findings overwhelmingly show that researchers improperly build the structure of their scales by using inappropriate statistics and methods such as principal components analysis, eigenvalues greater than one, and orthogonal rotation (Conway & Huffcutt, 2003; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Ford et al., 1986; Henson & Roberts, 2006; Norris & Lecavalier, 2010; Worthington & Whittaker, 2006). In the field of communication, three content analysis studies have been found examining the factor analytic practices of scholars finding questionable scale development procedures as well (Morrison, 2009; Park et al., 2002; Wimmer & Haynes, 1978). Furthermore, two studies found that scale development authors rarely included critical pieces of information for readers to evaluate the quality of scales (Morrison, 2009; Park et al., 2002).

Research questions

To verify the continued improper use of scale development procedures, a broad sample of communication journals was selected for a systematic quantitative content analysis. Scale development is a complex process that presents many options to scholars requiring several methodological and statistical competencies.

RQ1: To what extent will communication researchers apply improper scale development procedures in communication journals?

Second, the present research makes a unique contribution to literature by recording how authors identified dimensions and generated items for their proposed scales. Measurement model building practices should be evaluated in terms of not only methods and statistics, but concept explication best practices as well. As a result, choices prior to the application of statistical and methodological methods were reviewed to explore communication scholars' commitment to the development of valid measures.

RQ2: How do journal authors report how they generated and assessed dimensions and items for their proposed scales in communication journals?

Method

Descriptives

Articles that concentrated on survey research (79.0%) or experimental research (20.7%) were examined for this content analysis. Undergraduate students were targeted as respondents in a notable proportion (34.5%, $n = 207$) of the journal articles. If reported, response rates of surveys ranged from 5.0–99.0% with 6.3% ($n = 38$) of the studies reporting reaching less than a 25% response rate, 9.3% ($n = 56$) reporting between 26–50%, 8.0% ($n = 48$) reporting 51–75% response rate, and 4.8% ($n = 29$) with more than a 75% response rate. Notably, the majority of scholars (83.7%, $n = 502$) provided reliability levels for their scales and provided all scale items for readers (69.7%, $n = 418$). If reported, SPSS was the most used software used (7.0%, $n = 42$). In this study, the total explained variance by a scale was reported in 52.0% ($n = 312$) of the articles. The overall scale should account for the maximum amount of variance while not including items or dimensions that explain little variance. A scale should explain at least 50% of variance, but 75–90% is preferred (Beaver, et al., 2013; Streiner, 1994).

Sampling procedure

A quantitative content analysis of leading communication journals was selected for the purposes of describing the current state of scale development practices in the communication field. The 68 communication journals were identified based on the list of rankings from the Thomson Reuters' Journal Citation Reports. The unit of analysis was the journal article that included *exploratory factor analysis* or *principal components* to develop a latent measure. Authors of articles that included only a citation of a research study using factor analysis, only a confirmatory factor analysis using structural equation modeling, the use of factor analysis to develop a content analysis measure, or a mention, rather than the application, of the keyword search terms within the manuscript were removed from examination. This process resulted in a total of 1,318 journal articles from the 68 journals for the 10-year period.

A stratified random sampling procedure was employed because observations revealed that authors in the first 20 journals were more likely to employ scale development procedures. As a result, the articles were grouped based on their ranking: (1) 1–20, (2) 21–40, and (3) 41–68. A random sample of 600 articles (45.5% of the population) was selected: 300 articles from the top 20 journals, followed by 150 for the other 2 ranking categorizations from a total of 48 communication journals on the list. Several journals were not represented in the random selection process because the journals included no or very few studies using factor analysis.

Operational definitions

Scale development theoretical/methodological decision measures

Scale item and development measures included the reporting of relying on a literature review and/or theory, focus groups, interviews, cognitive interviews, q-sorts, pre-tests or pilot tests, experts, and other to create and assess items for scale development purposes. These variables were primarily assessed by examining the measures section within the method section. For example, literature review was coded as present if the authors cited other research that informed the creation of items for their proposed scale. All variables were treated as separate variables and dummy coded with *present* coded as 1 and *absent* as 0.

Major scale development analysis variables

The categories were based on previous content analyses and best practice recommendations from scale methodologists. The sample size used to conduct an EFA or PCA was broken down into categories including no report (see Table 2). The appropriateness of FA was examined by coding the presence or absence of Bartlett's test of sphericity and KMO. The methods of reduction included principal component analysis (PCA), common factor analysis (principal axis and/or maximum likelihood), multiple methods (PCA and common factor analysis), other, or no report. Oblique and orthogonal rotation method categories included oblique, orthogonal, both orthogonal and oblique, or no report. Rotation method type consisted of Promax, Direct Oblimin; Direct Quartimin, Varimax; Equamax; Quartimax, multiple rotation methods, other, and no report. The factor number determination options included eigenvalue greater than one, scree test, parallel analysis, minimum average partial, chi-square statistic, a priori number of factors retained (e.g., literature review suggested four factors), percentage of variance accounted for per factor, other approach, and no report. The scale item selection criteria categories included coding the presence and absence of minimal significant loadings (e.g., loadings above .40), cross loadings (i.e., significant loadings on two or more factors), item number per factor (e.g., minimum of 3 items per factor), inter-item correlations, theoretical convergence, communalities of variables, percentage of variance explained by a subscale, redundancy of wording or meaning across items, other item deletion criteria, and no report.

Table 2. Summary information of practices in the use of exploratory factor analysis in scale development ($n = 600$).

Characteristic	Frequency	%
Sample size*		
< 100	54	9.0
101–200	108	18.0
201–300	94	15.7
301–400	71	11.8
401–500	56	9.3
500 or larger	205	34.2
Ratio logic (e.g., 5:1; 10:1)	1	0.2
Factorability of data**		
Kaiser-Meyer Olkin (KMO)	11	8.9
Bartlett's test of sphericity	57	9.5
Type of analysis*		
Principal components analysis (PCA)	266	44.3
Common factor analysis	73	12.2
Multiple methods	13	2.2
Factor rotation method*		
Orthogonal	214	35.7
Varimax	205	34.2
Oblique	91	15.2
Direct Oblimin	27	4.5
Promax	24	4.0
Quartimax	3	0.5
Both orthogonal and oblique	8	1.3
Factor number determination criteria**		
Eigenvalue < 1	142	23.7
Scree test	45	7.5
A priori number of factors retained	20	3.3
Parallel analysis (PA)	9	1.9
Percentage of variance per factor	8	1.3
Minimum average partial (MAP)	2	0.3
Other number retention criteria	2	0.3
Item deletion or retention criteria**		
Factor loading magnitudes	154	25.7
Cross loadings	51	8.5
Inter-item correlations	35	5.8
Theoretical convergence	32	5.3
Factor number minimum	8	1.3
Percentage of variance	7	1.2
Communalities of variables	7	1.2
Item redundancy	6	1.1
Other criteria	7	1.2

Note. Variables do not add up to 100% because *no reports** were not included and some variables were absence/presence variables.**

Intercoder reliability

Reliability analyses of protocols are necessary when numbers represent the need to interpret meanings of the text. Riffe, Lacy, and Fico's (2014) sampling procedure was used to compute the test sample size for intercoder reliability resulting in 87 stories. These stories were randomly selected for intercoder reliability. Intercoder reliability on this sample was a challenge because of the nonexistent and minimal presence of some variables. The author of the study practiced coding articles not within the population to develop the protocol. It became clear that some variables would appear in a very small proportion of the articles. It was determined that these variables were still important due to the objectives of this study. Following the selection of articles, the PDF search function was employed to search for the presence of articles including the individual variables to add to the intercoder reliability analysis increasing the sample size to 114 for the author and one doctoral student. In addition, Riffe, Lacy, and Fico's (2014) recommend running multiple statistics to test the reliability of measures due to the intellectual debates associated with the appropriateness of certain reliability

statistics. I employed Krippendorff's Alpha, Cohen's Kappa, and Scott's Pi for reliability analyses for the nominal level variables. Reliability for the variables ranged from .79–1.0. For reliabilities of each variable and the codebook, please contact the author. Previous content analysis authors on factor analysis practices did not report intercoder reliability for their variables (Fabrigar et al., 1999; Henson & Roberts, 2006; Morrison, 2009; Norris & Lecavalier, 2010; Park et al., 2002; Worthington & Whittaker, 2006).

Results

The intent of data collection was to assess whether communication scholars have improved their scale development practices. RQ1 asked to what extent communication scholars followed scale development procedural best practices. As shown in Table 2, the present results showed that authors rarely reported that they inspected Kaiser-Meyer Olkin (KMO) (8.9%) or Bartlett's test of sphericity (9.5%) statistics prior to conducting factor analysis. Despite a consistent recommendation to not use PCA (e.g., Conway & Huffcutt, 2003; Costello & Osborne, 2005; Ford et al., 1986; Morrison, 2009; Norris & Lecavalier, 2010), 44.3% of the articles stated using it or they did not report the type of analysis used in 41.0% of the articles in this study. This PCA finding is proportionately less than in the Park et al. (52.9%) and more than in the Morrison (40.2%) studies on practices in communication journals.

In the present research, the eigenvalue greater than one rule (23.7%) was the most often applied method used to determine the number of factors or dimensions in a model (see Table 2), which is between the 27.7% found in the Park et al. (2002) study and 16.3% found in the Morrison (2009) study. The application of the eigenvalue rule was followed by the scree test (7.5%), theory/a priori number of factors (3.3%), and/or percentage of variance accounted by individual factors (1.3%).

In this present content analysis, orthogonal rotation (35.7%) was favored over oblique rotation (15.2%) despite recommendations to not use it. Specifically, communication scholars only reported applying the orthogonal rotation Varimax (34.2%). Notably, oblique rotation was applied slightly proportionately more often in comparison to the Park, Dailey, and Lemus (57.1% (orthogonal); 10.9% (oblique) and Morrison (34.2% (orthogonal); 11.4% (oblique) studies.

In the item deletion process, communication researchers most often relied on the rules associated with cross loadings and factor loading magnitudes. In the present study, however, more than 76.2% of the articles did not state whether authors relied on a priori cutoff criteria to determine item retention or deletion.

RQ2 queried what methods communication authors employed to capture the breadth of a concept prior to the launch of their quantitative study. The results showed that authors reported primarily relying on literature and/or theory to guide in their development of items (66.8%), followed by administering a pretest or pilot test (12.2%) with subjects. Authors, however, rarely reported seeking expert guidance (4.0%), administering interviews (2.0%), conducting focus groups (2.0%), conducting cognitive interviews (0.8%), using q-sorts (0.2%), or applying other approaches (3.5%).

10 steps in scale development and reporting

The multitude of choices involved in each scale development step have probably steered many researchers to shy away from best practices. The literature is sometimes technical, which may lead many users to simply trust the defaults in their statistical software packages. The goal is to highlight 10 major steps along the scale development decision tree to make the process more accessible and to encourage more systematic applications in future research.

STEP 1: research the intended meaning and breadth of the theoretical concept

Theory and research should play the strongest role in guiding the identification of empirical attributes that represent the abstract construct (Chaffee, 1991; Clark & Watson, 1995; Cronbach & Meehl, 1955; DeVellis, 2012). Theory should pre-specify the structure and meaning of a construct. The quality of a

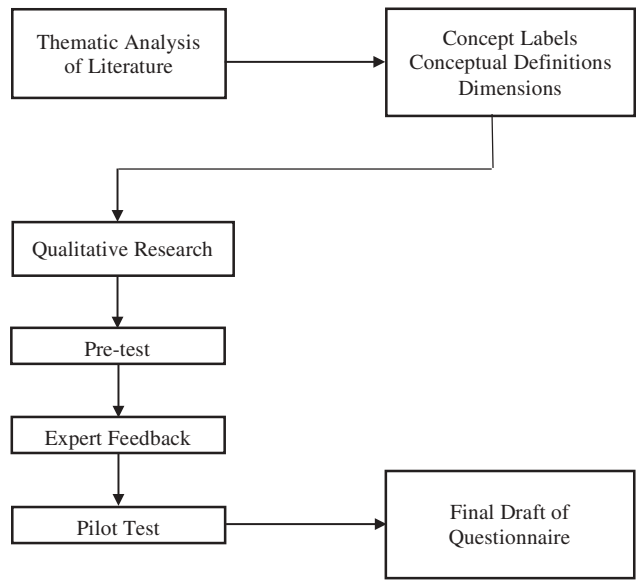


Figure 1. Research the intended meaning and breadth of the construct.

measure rests on the judgement of the researcher, which includes the selection of dimensions and wording of items to be included on the scale. Simms (2008) argued, “The apparent simplicity and efficiency of the [survey] method can be illusory, as much time, and consideration are needed to develop measures that allow us to make reliable and valid inferences about people (p. 414).” Ill-defined constructs require a large amount of time and effort on the part of the researcher to explicate the concept (Chaffee, 1991). The durability of measures would likely withstand statistical and methodological challenges if scholars relied on multiple methods to build them. Meaningful measurement occurs when the body of questions successfully achieves the intended representation of the abstract construct. Yet the specification of a theoretical concept can result in an indefinite number of items, but it is important to remember the goal is to find the optimal sample of items and dimensions to empirically represent its abstractness. Chaffee (1991) referred to this process as meaning analysis in which the theoretician employs logical procedures to justify the conceptual and operational definitions of the construct. Careful conceptualization, item selection, and wording are necessary to ensure content validity (Clark & Watson, 1995; Cronbach & Meehl, 1955; Worthington & Whittaker, 2006). The Step 1 section provides suggestions regarding how researchers should address the validity of the concept prior to the methodological applications and statistical analyses by addressing the labeling of the construct and dimensions, creating conceptual definitions, seeking to understand the breadth of the construct, and generating and refining items for the scale (see Figure 1).

Select appropriate conceptual labels

The naming of the construct and each subscale or dimension influences future interpretations of the concept. Researchers should be thoughtful when deciding what to label each concept. During a literature review, it is common to find conceptual redundancy even though concepts vary in labels within and across disciplines. As a result, a communication scholar should share the logic associated with the selection of a label. Unfortunately, scholars rarely shared the factor naming logic of their concepts in this study (3.5%, $n = 21$). One suggestion to address subjectivity is to invite a panel to review the items of each factor to determine the most appropriate concept label. Often times, the first items on each subscale can provide naming guidance for each dimension.

Select conceptual definitions

One beginning step in theory building is to write and defend a formal conceptual definition of the construct. A good definition adds clarity to an ambiguous concept by telling the reader what it is supposed to mean and what it is not supposed to mean (Chaffee, 1991). For example, *quality of life* could refer to attributes such as mobility, quality of sleep, eating problems, but it also could include water quality, noise levels, and pollutants. And, thus, we have to be conceptually clear by providing a dictionary-like definition with concrete keywords or attributes that are useful for academic audiences, which then should influence the framing and wording of items included the scale.

Identify potential dimensions and items

The literature review should not only be conducted through the lens of the construct label, but authors should seek literature to identify possible dimensions of the construct and defend their association with the overarching construct. An additional function of the literature review is to gain a more concrete understanding of the construct by identifying conceptual definitions of each dimension and relevant empirical items that describe each dimension.

Conduct qualitative research to generate and validate dimensions and items

Several rounds of information gathering should be conducted following the conceptualization of the construct and prior to full data collection to reduce measurement error. The pool of scale items needs to be concise, clear, distinct, and reflect the chosen conceptual definition (DeVellis, 2012). Interviews, focus groups, and expert feedback are critical in the item generation and dimension identification process (Broom, 2006; Clark & Watson, 1995; DeVellis, 2012; Pett et al., 2003; Simms, 2008; Worthington & Whittaker, 2006). Focus group and interview research, however, does not appear often in media and communication journals in the United States (Potter & Riddle, 2007; Ye & Ki, 2012). Furthermore, previous content analyses of communication journals found that scholars very rarely combined quantitative and qualitative methods (Cooper, Potter, & Dupagne, 1994; Kamhawi, 2003; Trumbo, 2004; Ye & Ki, 2012).

This is unfortunate, because based on experience, it is very common that participants will reveal additional dimensions critical to the meaning of the construct based on results stemming from qualitative research efforts. In regard to developing a qualitative research protocol with the specific goal of scale development in mind, researchers should begin by asking participants broadly about their interpretation of the construct, followed by questions concerning the participants' interpretations of each dimension found in the literature review; including to what degree they agree with each dimension. The researcher should probe the participants for possible scale items including the wording of items representing each dimension as well.

Use feedback to refine scale

Measurement error can arise for many reasons such as complex wording or language, questions requiring estimation, vagueness in questions or response categories, double-barreled questions, and leading or biased questions. Q-sorts, pilot tests, expert feedback, cognitive interviews, and pretests are especially useful in questionnaire and item refinement (Clark & Watson, 1995; DeVellis, 2012; Pett et al., 2003; Worthington & Whittaker, 2006).

Pre-test. Pre-tests on smaller samples are useful for survey and item feedback prior to the launch of the data collection, while pilot tests, following a pre-test, can be employed to assess how the data will fall to determine whether items should be added or deleted. In this study, scholars struggled with distinguishing a pre-test from a pilot test, which resulted in the combining of both categories into one for coding purposes. The use of pre-tests in surveys dates back to the 1930s. Researchers can employ multiple pre-tests to refine their questionnaire questions and design. Pretesting can address areas such as ambiguous, leading, confusing, difficult, skipped, sensitive, and missing questions. The

goal is to reduce measurement error, response burden, and question inaccuracy. Pretest sample sizes should be small, but similar as possible to targeted respondents. Pretest sample sizes can range from 5–100 people depending upon the diversity of target subpopulations.

Pretesting can be conducted with focus groups, cognitive interviews, interviews, group debriefing, or individual debriefing. Cognitive interviews consist of probes (e.g., “What does ‘some of the time’ mean to you?” and think-alouds (i.e., the interviewer requests that the respondent reads each question and verbalizes what comes to mind when reading each question.). Scholars can conduct behavioral coding such as watching whether respondents hesitate or frown when reading a question. Behavioral coding consists of monitoring or reviewing taped interviews or survey participation. Following one or multiple pre-tests, statements, instructions, and the questionnaire design should be edited based on this feedback (Couper, Lessler, Martin, Martin, Rothgeb, & Singer, 2004; Drennan, 2003; Lewis, Templeton, & Bryd, 2005; Reynolds, Diamantopoulos, & Schlegelmilch, 1993; Ruel, Wagner, & Gillespie, 2016).

Expert feedback. Experts should consist of methodologists, intended participants, and subject-matter researchers. The goal is get their feedback on item quality and how well each item reflects the overarching construct. RESEARCHERS can provide instruction to the experts BY asking them TO provide individual feedback on items BY asking them to assess item validity through a Likert-type scale OR open-ended feedback (DeVellis, 2012; Ruel et al., 2016).

Pilot test. A pilot test is rehearsal of the actual survey in actual field conditions. The quantitative data collection part of a pilot test is especially useful in identifying how data will fall around each factor and identifying skipped questions. In order to conduct an EFA, the pilot test sample size should range from 50–100 participants. Once edits based on the pilot test are complete, the survey is set for full-scale administration (Lewis, Templeton, & Bryd, 2005; Ruel et al., 2016). At a minimum level, scholars should employ: (1) a literature review; (2) at least one type of qualitative research; (3) expert feedback; and (4) a pre-test when developing their scale dimensions and items.

STEP 2: determine sampling procedure

Decide an appropriate sample size

Following the content development stage, scholars can proceed toward factor analysis. Factor analysis is a large sample size technique. Insufficient sample sizes result in unstable factors and decreased generalizability (Kline, 2013; Tabachnick & Fidell, 2007). Methodologists vary regarding recommended sample sizes with the exception of stating that more participants result in more stable scales. Generally, most scholars recommend a sample size of at least 300 (McCroskey & Young, 1979; Henson & Roberts, 2006; Pett et al., 2003; Worthington & Whittaker, 2006). Recommendations range from a sample size of 50 (Barrett & Kline, 1981)–400 (Aleamoni, 1976). Comrey and Lee (1992) provided one guide: 50 (very poor), 100 (poor), 200 (fair), 300 (good), 500 (very good), and 1000 (excellent). If the communalities and factor loadings are low, it is suggested to increase the sample size (Mundfrom, Shaw, & Ke, 2005). A communality (h^2) is the proportion of variance accounted by each individual variable for one factor. In this study, the communalities for each item were shared in only nine studies. Communalities are considered high if they are above .80; however, the more common range in the social sciences is from .40–.70 (Costello & Osborne, 2005). Thus, lower sample sizes can be defended if a majority of communalities (<.50) and factor loadings (<.40) are high (see Worthington & Whittaker, 2006).

A scale with fewer variables, however, requires fewer participants. Every scale varies in dimensions, item numbers, communality sizes, factor correlations, and item-factor correlations (Floyd & Widaman, 1995; Osborne, 2014; Pett et al., 2003; Worthington & Whittaker, 2006). For these reasons, Guadagnoli and Velicer (1988) and other methodologists argue that item ratios are more relevant than the previously mentioned sample size defense logics. Gorsuch (1983) and others

suggested following lower minimum ratios of participants to items (5:1 or 10:1), while Osborne (2014) argued for 20 cases per variable to ensure robust, generalizable results. Costello and Osborne (2005) addressed the sample size debate by examining how varying sample sizes affected error rates regarding the factor structure of a scale. Larger solutions (20:1) produced the most correct solutions and classification of items.

Many studies with smaller sample sizes (e.g., 100 people) are published (Conway & Huffcutt, 2003; Russell, 2002). It is very common to use adequate rather than ideal sample sizes especially when dealing with difficult to access populations (Worthington & Whittaker, 2006). For example, one content analysis on a subset of public relations journals showed that a notable proportion (42.8%) of articles published relied on a sample size ranging from 101–200 people (Ki & Shin, 2006). Approximately 43% of the journal articles in this study included a sample size of less than 300 (see Table 2). Based on coder observations, studies with lower sample sizes surveyed professionals (e.g., journalists, public relations practitioners) or utilized the experimental method. Lower sample sized studies are quite common based on previous content analysis results (Conway & Huffcutt, 2003; Fabrigar et al., 1999; Henson & Roberts, 2006).

Recommendations for a sufficient sample size are challenging for these reasons. Additionally, methodologists recommend using two-to-three times as many items than one expects to be on the final scale. For example, a communication scholar would ideally write 60 items that would ultimately result in a final 20-itemed scale, which means the sample size would need to consist of 1200 participants to follow the 20:1 rule of thumb. There is no clear consensus. To be of use beyond a particular sample, the most optimal recommendation to follow is the 20:1 ratio logic to reduce the error rate, but communication scholars should abide by the minimum standard of a 5:1 item ratio of participants to number of variables.

STEP 3: examine data quality

Data cleaning is essential to ensure that findings are accurate and replicable. Researchers should check for missing data, absence of outliers, linearity, and extreme multicollinearity (Beavers et al., 2013). For example, outliers should be justifiably changed or deleted if only a few exist because of their impact on outcomes. Missing data are not ignorable, the researcher should inspect patterns of missing data. Scholars should consider deleting cases when the majority of responses (50% or more) contain missing data. It is encouraged to read books on data cleaning best practices (e.g., Hair, Black, Babin, & Anderson, 2010; Meyers, Gamst, & Guarino, 2006; Myers, 2011). One suggestion is to communicate to readers in research articles how one dealt with missing data, outliers, or other potential issues.

STEP 4: verify the factorability of the data

Inspection of the correlation matrix, Bartlett's test of sphericity, and Kaiser-Meyer-Olkin (KMO) provides information as to whether factor analysis should be applied to data. The correlation matrix should include numbers at levels of .30 or higher, Bartlett's chi-square should be significant at a probability of .05 or less, and a KMO value of .60 or higher is recommended before proceeding with factor analysis (McCroskey & Young, 1979; Pett et al., 2003; Tabachnick & Fidell, 2007). Previous research, however, showed that authors do not often report that they examined these statistics prior to proceeding with a factor analysis (Henson & Roberts, 2006; Worthington & Whittaker, 2006).

STEP 5: conduct common factor analysis

The group of extraction and rotation techniques to identify latent constructs is referred to as *common factor analysis* or *exploratory factor analysis*. It is important to be aware that factor analysis is not one statistical procedure, but rather a group different statistical and methodological choices

(Beavers et al., 2013). And thus, the accuracy of the results rests on the quality of these decisions made during each step.

EFA seeks to identify the shared variance among variables. One common mistake made by researchers is the use of principal components analysis (PCA) (Goldberg & Velicer, 2006; Reise et al., 2000). PCA is conceptually and mathematically distinct from common factor analysis. Common factor analysis does not focus on explaining the amount of variance, but rather it assesses the sources of common variation. The problem with PCA in latent measurement development is that the sizes of components are inflated because they include error variance, which can lead researchers to retain too many components (Conway & Huffcutt, 2003; Costello & Osborne, 2005; Goldberg & Velicer, 2006; Preacher & MacCallum, 2003; Snook & Gorsuch, 1989). Common factor analysis (i.e., principal axis factoring or maximum likelihood) results are more generalizable when submitting hypothesized models to a confirmatory factor analysis (Haig, 2005; Worthington & Whittaker, 2006). And, thus, it is recommended to conduct common factor analysis rather than PCA because most methodologists do not support the application of it for the previously mentioned reasons.

STEP 6: select factor extraction method

The extraction method evaluates the correlation/covariation among all of the scale items and seeks to extract latent variables from the manifest variables (Osborne, 2014). Essentially, we are interested in factors, and we are transforming our data from a variable space to a factor space. Statistical package options include unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring, and image factoring. Osborne (2014) noted little information exists on the drawbacks and advantages of each method. Based on observations and research, scholars most often select either principal axis factoring (PAF) or maximum likelihood (ML). Both MA and PAF try to reproduce the correlation matrix. PAF, the most robust method, is recommended to use when sample sizes are small and normality is violated, and ML is supported when data is normally distributed (Fabrigar et al., 1999; Nunnally & Bernstein, 1994).

STEP 7: determine number of factors

Next, the precise number of the subscales to include is rarely clear cut. I will review the four common approaches used to determine factor number.

Eigenvalue greater than one

The most often used, but not supported approach, is the eigenvalues greater than one rule when assessing factor number (Ford et al., 1986; Henson & Roberts, 2006; Morrison, 2009; Russell, 2002). Eigenvalues measure the variance accounted for by each factor. The larger the eigenvalue, the more variance explained by a factor. In the past, Kaiser (1960) believed that eigenvalues greater than one resulted in stable dimensions. Today, several methodologists advise against using the cutoff of one eigenvalue when determining the number of factors because the number of factors above one is greatly related to the total number of variables included a model, which researchers argue leads to the extraction of too many or too few factors (Fabrigar et al., 1999; Goldberg & Velicer, 2006; Kline, 2013; Worthington & Whittaker, 2006). And, thus, it is not recommended for the development of measurement models.

Scree test

The scree test is a graphic that allows researchers to estimate the number of factors to retain. A scree is a visual plot of eigenvalues derived from the factors. The cutoff line for number of factors is determined when a line elbows off from a somewhat subjectively straight dotted line. The scree test is considered more accurate than the eigenvalue rule (McCroskey & Young, 1979; Pett et al., 2003; Preacher & MacCallum, 2003; Reise et al., 2000).

Parallel analysis (PA)

PA was developed by Horn (1965), and it compares eigenvalues from the results against a randomly ordered data set. The PA method has been recommended for determining an accurate number of factors to accept (Humphreys & Montanelli, 1975; Kline, 2013; Velicer, Eaton, & Fava, 2000; Watkins, 2006; Zwick & Velicer, 1982). Scholars should compare their data's eigenvalues with the eigenvalues produced in PA. Factors are retained when their eigenvalues are larger than the eigenvalues created by a random data set. A lack of awareness and the need to download the free program may be reasons for people not adopting PA. However, journals such as *Educational & Psychological Measurement* and the *Journal of Personality Assessment* require reporting of PA in scale development articles (Pallant, 2010). To access the stand-alone program, I recommend visiting <http://edpsychassociates.com/Watkins3.html> for a free download of Monte Carlo PCA for Parallel Analysis (Watkins, 2006).

Minimum average partial (MAP)

MAP involves partialling each successive factor from a correlation matrix to create a partial correlation matrix. MAP, introduced by Velicer, examines off-diagonal partial correlations after successively removing the effects of factors (Velicer, 1976). The average of the squared correlations of the off-diagonal partial correlation matrix is computed after each factor is extracted and its effect on the correlations between items is excluded. The average should decline as long as shared variance is being extracted, and the average will then increase when error variance dominates. Factors are retained when the averaged square partial correlation reaches its lowest value (Goldberg & Velicer, 2006; Watkins, 2006). MAP is not as readily available, which likely partially explains its low use. O'Conner (2000) provided guidance on HOW TO USE these procedures WITH SPSS, SAS, and MATLAB syntax commands for both MAP and PA. (<https://people.ok.ubc.ca/briocconn/nfactors/nfactors.html>).

Social scientists concerned about the optimal number of factors should determine them based on theory and multiple tools. Researchers should use a combination of the following: theory/literature review, visual scree test, parallel analysis, and minimum average partial when determining the appropriate number of factors in a hypothetical model (Conway & Huffcutt, 2003; DeVellis, 2012). If these previously mentioned procedures reveal that multiple factor numbers are a possibility, such as 3, 4, and 5 factor models, authors should rerun analyses examining 2–6 factor models to determine the optimal factor number for the scale.

STEP 8: rotate factors

Rotation is necessary in order to more clearly identify the scale's factors (or dimensions). Oblique and orthogonal are two types of rotation methods available to researchers. Content analyses reveal most scholars use Varimax, an orthogonal rotation, which forces factors to not correlate (Borgotta, Kercher, & Stull, 1986; Conway & Huffcutt, 2003; Ford et al., 1986; Norris & Lecavalier, 2010).

Orthogonal produces uncorrelated factors. The problem is that it is uncommon for factors to not correlate with one another in the social sciences (DeVellis, 2012; McCrosky & Young, 1979). A Varimax rotation, the most popular orthogonal rotation type, pushes high factor loadings higher and low factors lower because they are not allowed to correlate (Tabachnick & Fidell, 2001). Varimax is biased against finding a general factor when one exists and it generates more cross loadings (Fabrigar et al., 1999; Gorsuch, 1997). As a secondary analysis of data shows in Table 3, a Varimax rotation resulted in 9 significant cross loadings. It is also produced 13 lower and 12 higher loadings than a Promax rotation. If factors are substantially correlated, an oblique solution improves the ability of the researchers to approximate the structure of the model (Fabrigar et al., 1999; Ford et al., 1986; Gorsuch, 1997). If factors do not correlate, scholars should assess whether they are measuring two separate constructs rather than one. Thus, it is recommended to use oblique rotation because it more accurately represents most models in communication research because it allows factors to correlate.

Table 3. Secondary data analysis comparison of orthogonal (1st item) and oblique (2nd item) rotations ($n = 551$).

Items	1	2	3	4	5	6	7
1	0.53	-0.02	0.03	0.37	-0.16	0.06	0.08
1	-0.05	-0.02	0.69	-0.03	-0.06	0.12	0.05
2	0.46	0.27	0.06	-0.12	0.07	0.02	-0.09
2	0.36	0.13	-0.09	-0.15	0.21	0.10	0.08
3	0.26	-0.21	0.03	0.22	-0.13	0.07	0.12
3	-0.06	0.00	0.46	0.08	-0.20	0.00	0.02
4	0.39	0.32	0.07	0.12	0.19	-0.27	-0.15
4	0.04	-0.04	0.03	0.07	0.60	0.11	-0.01
5	0.45	0.04	0.14	0.48	-0.07	0.03	-0.01
5	-0.09	-0.20	0.76	0.02	0.12	0.09	0.14
6	0.54	0.12	0.04	0.31	0.06	-0.02	0.04
6	-0.04	0.10	0.51	-0.03	0.20	0.03	0.07
7	0.51	0.26	0.10	0.27	0.23	-0.21	0.08
7	-0.01	0.10	0.35	0.01	0.48	-0.12	-0.04
8	0.24	0.31	0.05	-0.05	0.15	-0.20	-0.15
8	0.14	-0.02	-0.18	-0.01	0.47	0.10	-0.03
9	0.55	0.08	-0.24	-0.02	0.04	0.00	0.12
9	-0.03	0.55	0.08	-0.11	0.02	0.07	-0.09
10	0.45	0.10	0.05	0.09	0.13	0.04	0.18
10	0.14	0.29	0.27	-0.11	0.05	-0.19	0.01
11	0.62	0.08	-0.36	-0.02	0.10	0.08	-0.07
11	-0.14	0.67	-0.02	-0.09	0.07	0.25	0.07
12	0.60	-0.05	-0.44	-0.11	0.10	0.10	0.07
12	-0.17	0.88	-0.08	-0.08	-0.10	0.15	-0.01
13	0.56	-0.08	-0.11	-0.08	-0.25	-0.18	0.25
13	0.18	0.22	0.13	0.05	-0.12	0.12	-0.37
14	0.57	0.05	-0.03	-0.06	-0.03	-0.11	0.19
14	0.23	0.28	0.09	-0.01	0.05	-0.01	-0.20
15	0.44	0.11	0.33	-0.24	0.04	0.11	0.03
15	0.76	-0.04	-0.08	-0.08	0.01	-0.13	0.10
16	0.44	-0.14	-0.07	-0.02	0.23	0.11	0.03
16	0.02	0.49	0.04	0.09	0.00	-0.09	0.16
17	-0.06	0.50	-0.05	-0.03	-0.02	0.18	0.14
17	0.13	0.12	0.04	-0.61	-0.06	-0.11	0.00
18	0.20	-0.65	0.11	0.05	0.02	-0.14	-0.06
18	-0.06	-0.07	0.08	0.70	-0.04	0.03	0.02
19	0.46	-0.08	0.07	-0.09	0.19	0.11	-0.12
19	0.25	0.25	-0.06	0.12	0.09	0.01	0.23
20	0.45	-0.62	0.11	-0.02	0.18	-0.15	0.00
20	0.05	0.21	0.01	0.71	0.05	-0.09	0.03
21	0.57	-0.15	0.03	0.03	0.05	0.28	-0.13
21	0.21	0.26	0.20	0.05	-0.12	0.12	0.33
22	0.56	-0.24	0.14	-0.15	-0.09	-0.06	0.01
22	0.43	0.05	0.00	0.27	-0.05	0.09	-0.05
23	0.50	-0.19	-0.17	-0.17	0.26	-0.02	0.11
23	0.04	0.69	-0.19	0.17	0.03	-0.12	-0.03
24	0.67	0.05	-0.24	-0.09	-0.22	0.03	-0.14
24	0.13	0.28	0.00	-0.05	-0.03	0.49	-0.03
25	0.59	-0.09	0.09	-0.17	-0.13	0.05	0.01
25	0.49	0.10	0.01	0.07	-0.10	0.14	-0.01
26	0.62	0.00	-0.02	0.10	-0.33	-0.03	-0.16
26	0.18	-0.14	0.29	0.06	0.00	0.51	-0.01
27	0.59	0.04	-0.33	0.04	-0.14	-0.13	-0.23
27	-0.17	0.26	0.00	0.09	0.19	0.58	-0.05
28	0.54	0.16	0.21	-0.27	-0.07	-0.03	0.00
28	0.71	-0.03	-0.15	-0.06	0.09	0.07	-0.06
29	0.47	0.08	0.38	-0.20	-0.07	-0.04	-0.03
29	0.77	-0.26	-0.05	0.06	0.10	0.00	0.00
30	0.52	0.15	0.19	-0.05	-0.06	-0.06	0.07
30	0.47	-0.04	0.12	-0.06	0.11	0.00	-0.07
31	0.48	0.10	0.16	-0.11	-0.04	0.17	0.03
31	0.51	0.08	0.11	-0.16	-0.09	0.00	0.11
32	0.32	-0.11	0.18	0.23	0.11	0.30	-0.08
32	0.10	0.02	0.45	0.02	-0.09	-0.06	0.40

In an oblique rotation, researchers can review two matrices, but the pattern matrix should be evaluated to assess salience and simple structure. Oblique rotation options include Direct Oblimin and Promax. Both allow factors to correlate, but Promax begins with an orthogonal solution and then transforms it to an oblique solution (Hendrickson & White, 1964). Promax has been argued to be more robust, and it is recommended (Thompson, 2004).

STEP 9: retain and delete items based on a priori criteria

Ideally, scholars include around three times as many items on a questionnaire than will be included in the final scale. Following the item writing process, researchers need to winnow down the number of items by deciding which items to retain or discard. Items can be referred to as variables and questions as well. It is critical to optimize scale length to ensure participant motivation.

Simple factor structure is often determined based on several pre-established general criteria: factor item loadings at or above the .30–.50 level, no cross-loadings (i.e., significant loadings on more than one factor), no factors with fewer than three items, reliability levels, and theoretical convergence (Clark & Watson, 1995; Costello & Osborne, 2005; DeVellis, 2012; Fabrigar et al., 1999; Gorsuch, 1997; Hair et al., 2010; Kline, 2013; Norris & Lecavalier, 2010; Tabachnick & Fidell, 2007; Tinsley & Tinsley, 1987; Worthington & Whittaker, 2006). If reported in journal articles, cross loadings and significant loading magnitude levels are the most commonly applied criteria (Worthington & Whittaker, 2006).

It is suggested that scholars evaluate items based on multiple criteria: theory, communalities, items loadings, no significant cross-loadings, minimum of three salient loadings, factor reliability levels, and parsimony. I will overview minimum item loadings and number of items per factor because these issues appeared most often in the results.

Minimum factor item loading

In the present study of communication research, authors tended to vary in their factor loading cutoff levels if they were reported: (1) .20 or lower (0.3%) and (2) .30–.39 (3.7%), .40–.49 (7.5%), .50–.59 (6.7%), .60–.69 (5.3%), .70 or higher (0.3%). A factor loading is a correlation between an item and a factor. Loadings can be positive or negative depending on their correlation with other variables. Based on coder observations, it appears that several scholars followed the 60–40 criterion developed by McCroskey and Young (1979) to determine *significance*. That rule, however, is not supported by outside scholars. Based on factor analysis recommendations, acceptable levels are notably lower: .30 (Kachigan, 1986; Pett et al., 2003; Russell, 2002; Tinsley & Tinsley, 1987), .32 (Worthington & Whittaker, 2006), .35 (Clark & Watson, 1995), .40 (Ford et al., 1986; Hair et al., 2010; Reinard, 2006), and .50 (Mertler & Vannatta, 2001). A squared correlation of .30 equals .09 and a .32 means the item accounts for 10.2% of the overlapping variance with the items within a factor. Based on a review of recommendations and due to the variations found in the present study, it is recommended that a significant cut-off level be at .32, but anywhere between .30–.40 is supported by the literature review.

Number of items per factor

It is recommended that each subscale include at least three items in order to capture the true central of each dimension. Methodologists recommend being over-inclusive with the number of items (Clark & Watson, 1995; Kline, 2013; Loevinger, 1957). One flaw with many scales is that they use one- or two-item measures to tap into an abstract concept. Two-itemed scales are only recommended if items are highly correlated (i.e., $r < .70$) (Worthington & Whittaker, 2006). An additional analysis of the articles in this study found that 15.8% of the journal articles included two-itemed factors in their new scales. Most methodologists endorse that each factor should include a minimum of three variables, however, at least four or five variables per dimension are recommended (Costello & Osborne, 2005; Fabrigar et al., 1999; Gorsuch, 1983; Kline, 2013; Reise, Thurstone, 1947; Waller & Comrey, 2000).

It is suspected, however, that many scales contain highly redundant items to achieve this ideal of three items per factor. During the development stage, scholars can select a few empirical items that are redundant in meaning in order to identify the best statements that represent the construct. The final scale, however, should not contain insufficiently distinct items that inflate reliability levels and have a negative impact on the goal of parsimony. In fact, high coefficient alpha levels may suggest an over-inclusion of certain items. Qualitative research is key to preventing redundancy issues. Scale items should be related, but also be distinct aspects of the latent factor. This goal is very difficult to accomplish, but it is a hallmark of the best scales.

STEP 10: present results

An important point that needs to be addressed is the continued practice of not reporting information regarding the logic and choices made at major decision points in the scale development process. Scholars should report the following information: construct and subscale naming logic and conceptual definitions, sample size logic, methods for determining factor numbers, Bartlett's test of sphericity and Kaiser-Meyer-Olkin test of sampling adequacy results, factor extraction method, rotational method, strategies for selecting items, eigenvalues for all factors, pattern matrix, computer program package, communalities for each variable, descriptive statistics, subscale reliabilities, and percentage of variance accounted for by each factor.

The discussion of results stemming from a scale development manuscript should reflect a discussion explaining each factor including the naming logic and conceptual definitions associated with any new factors. Scale development research explores factors, and it does not present hypotheses and research questions (Pett et al., 2003).

Conclusion

My intent is to empower the reader by dissecting the fundamentals of the scale development methodology in hopes of advancing communication and media research because scale development is a complex, multi-step process. Scale development is truly a craft that requires practice. Good theorists should explicate their concepts by precisely defining them. Thoughtful attention to measurement may result in a forced clarification of our concepts that represent our fields, which will then enable us to stand stronger on our body of scientific knowledge—what we claim we know.

The misuse associated with exploratory factor analysis likely stems from a lack of awareness in the field of communication. As noted, the reliance on statistical package defaults most often goes against best practices in scale development. The accuracy and soundness of our predictions rests heavily on not only our statistical and methodological decisions, but our theoretical logic as well. This article makes a unique contribution by studying and reviewing the concept explication process involved with scale development. It is important to note that the measurement model building process is exploratory, and a resulting measure may continue to evolve over time. A scientist should not hesitate to move back and forth between meaning and empirical analysis until the precision and structure of a measure adequately represents the latent construct because it is not a linear process (Chaffee, 1991; Osborne, 2014).

Educational leaders including professional associations can address these issues. The hiring of methodologists at educational institutions, the support of methodological and theoretical divisions, and the teaching of scale development can influence graduate students and faculty members to enact a more critical eye toward measurement quality. To engage students, it is important to conceptually explain the statistical and methodological decision logics in order to encourage the adoption of best practices. In the meantime, many reputable and dedicated methodologists have written guidelines on the ideal statistical and methodological choices associated with scale development, and I encourage readers to read the literature listed in the references section for additional guidance or read some scale development papers for manuscript writing guidance (Carpenter, Grant, & Hoag, 2016; Carpenter, Makhadmeh, & Thornton, 2015).

Future research needs to be conducted to encourage measurement literacy in other areas. For example, a content analysis could focus on index development practices in communication. Methodologists refer to indices and summated scales as two different types of measures. An index is a formative measure, whereas a summated scale is a reflective measure. Some example differences between the two include that the deletion of one item from an index negatively affects the theoretical meaning of a construct, and indices, made up of manifest items such as what is often found in content analysis work, are not expected to correlate with one another (Bollen & Lennox, 1991; Morrison, 2009). In the present study, it was found that several scholars referred to a summated scale as an index (11.7%). As a result, it will be challenging to content analyze index development efforts until the communication field addresses the differences between the two types of measures. A future content analysis could also assess how scholars applied confirmatory factor analysis techniques to validate scales. Several content analyses have been done outside the field of communication evaluating such practices (Jackson, Gillaspay, & Pure-Stephenson, 2009; Schreiber et al., 2006; Worthington & Whittaker, 2006). This article does not cover all measurement approaches. And thus, it is hoped that this article will simply offer readers a foundation on measurement theory and procedures.

Researchers could conduct a content analysis of manuscript reviews to determine how often factor analysis is mentioned in reviews and examine what type of guidance is provided by editors and reviewers. Additionally, a comparative analysis of journals with lower or higher word limits could be carried out to determine the extent in which manuscript word limits play a part in the communication of measurement development practices.

Journal editors and reviewers play a critical role in raising the field to these standards. I would encourage journal editors to welcome manuscripts that focus on measurement development. Not all science consists of hypothesis testing; science can also consist of the theoretical development of measures. Most of the articles in this study concentrated on both scale development and hypothesis testing (92%) rather than scale development (8%). The targeted attentiveness on measurement development by journal editors may help clarify concepts and teach journal readers about best practices.

Acknowledgments

I would like to thank Marley Watkins who provided me with a strong foundation in scale development and inspired me to continue down this path in the field of communication. I would also like to acknowledge the editor Jörg Matthes and the reviewers for encouraging my work to be much better.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

ORCID

Serena Carpenter  <http://orcid.org/0000-0002-8814-4185>

References

- Aleamoni, L. M. (1976). The relation of sample size to the number of variables in using factor analysis techniques. *Educational and Psychological Measurement*, 36, 879–883.
- Barrett, P. T., & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality Study and Group Behavior*, 1, 23–33.
- Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in education research. *Practical Assessment, Research & Evaluation*, 18(6), 1–13.
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement – a structural equation perspective. *MIS Quarterly*, 28(2), 305–314.

- Borgotta, E. F., Kercher, K., & Stull, D. E. (1986). A cautionary note on use of principal components analysis. *Sociological Methods and Research*, 15, 160–168.
- Broom, G. M. (2006). An open-system approach to building theory in public relations. *Journal of Public Relations Research*, 18(2), 141–150.
- Carpenter, S., Grant, A. E., & Hoag, A. (2016). Journalism Degree Motivations (JDM): The development of a scale. *Journalism & Mass Communication Educator*, 71(1), 5–27.
- Carpenter, S., Makhadmeh, N., & Thornton, L. J. (2015). Mentorship on the doctoral I level: An examination of communication mentors' traits and functions. *Communication Education*, 64(3), 366–384.
- Chaffee, S. H. (1991). *Explication* (pp. 1–42). Beverly Hills, CA: Sage.
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309–319.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, 6(2), 147–168.
- Cooper, R., Potter, W. J., & Dupagne, M. (1994). A status report on methods used in mass communication research. *Journalism Educator*, 48(4), 54–61.
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7), 1–9.
- Couper, Mick P., J. Rothgeb, J. Lessler, E.A. Martin, J. Martin, and Eleanor Singer. (2004). *Methods for Testing and Evaluating Survey Questionnaires*. New York: Wiley.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological test. *Psychological Bulletin*, 52, 281–302.
- DeVellis, R. F. (2012). *Scale development. Theory and applications* (3rd ed.). Thousand Oaks, CA: Sage.
- Drennan, J. (2003). Cognitive interviewing: Verbal data in the design and pretesting of questionnaires. *Journal of Advancing Nursing*, 42(1), 57–63.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272–299.
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7(1), 286–299.
- Ford, J. K., MacCullum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, 39, 291–314.
- Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103, 265–275.
- Goldberg, L. R., & Velicer, W. F. (2006). Principles of exploratory factor analysis. In S. S. Strack (Ed.), *Differentiating normal and abnormal personality* (pp. 209–237). New York, NY, USA: Springer.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, 68, 532–560.
- Haig, B. D. (2005). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research*, 40(3), 303–329.
- Hair, J., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Mathematical Psychology*, 17, 65–70.
- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research. Common errors and some comment on improved practice. *Education and Psychological Measurement*, 66(3), 393–416.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185.
- Humphreys, L. G., & Montanelli, R. G., Jr. (1975). An investigation of the parallel criterion for determining the number of common factors. *Multivariate Behavioral Research*, 10, 193–205.
- Jackson, D. L., Gillasp, J. A., Jr., & Pure-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, 14(1), 6–23.
- Kachigan, S. K. (1986). *Statistical analysis: An interdisciplinary introduction to univariate and multivariate methods*. New York: Radius Press.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141–151.
- Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika*, 35(4), 401–415.
- Kamhawi, R. (2003). Mass communication research trends from 1980 to 1999. *Journalism & Mass Communication Quarterly*, 80(1), 7–27.
- Ki, E., & Shin, J. (2006). Status of organization-public relationship research from an analysis of published articles, 1985–2004. *Public Relations Review*, 32, 194–195.
- Kline, R. B. (2013). Exploratory and confirmatory factor analysis. In Y. Petscher, C. Schatschneider, & D. L. Compton (Eds.), *Applied quantitative analysis education and the social sciences* (pp. 171–207). New York, NY, USA: Routledge.

- Levine, T. R., Hullett, C. R., Turner, M. M., & Lapinski, M. K. (2006). The deirability of using confirmatory factor analysis on published scales. *Communication Research Reports*, 23, 309–314.
- Lewis, B. R., Templeton, G. F., & Byrd, T. A. (2005). A methodology for construct development in MIS research. *European Journal of Information Systems*, 14(4), 388–400.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(7), 635–694.
- McCrosky, J. C., & Young, T. J. (1979). The use and abuse of factor analysis in communication research. *Human Communication Research*, 5, 375–82.
- Mertler, C. A., & Vannatta, R. A. (2001). *Advanced and multivariate statistical methods: Practical applications and interpretation*. Los Angeles, CA: Pyrczak Publishing.
- Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation*. Thousand Oaks, CA: Sage.
- Morrison, J. T. (2009). Evaluating factor analysis decisions for scale design in communication research. *Communication Methods and Measures*, 3(4), 195–215.
- Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, 5, 159–168.
- Myers, T. A. (2011). Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures*, 5(4), 297–310.
- Norris, M., & Lecavalier, L. (2010). Evaluating the use of exploratory factor analysis in developmental disability psychological research. *Journal of Autism Development and Disorders*, 40, 8–20.
- Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York, NY, USA: McGraw Hill.
- O'Conner, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Veliver's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32(3), 396–402.
- Osborne, J. W. (2014). *Best practices in exploratory factor analysis*. Scotts Valley, CA: CreateSpace Independent Publishing.
- Pallant, J. (2010). *SPSS: Survival manual*. New York, NY, USA: McGraw Hill.
- Park, H. S., Dailey, R., & Lemus, D. (2002). The use of exploratory factor analysis and principal components analysis in communication research. *Human Communication Research*, 28(4), 562–577.
- Pett, M. A., Lackey, N. R., & Sullivan, J. L. (2003). *Making sense of factor analysis. The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage Publications, Inc.
- Potter, W. J., & Riddle, K. (2007). A content analysis of the media effects literature. *Journalism & Mass Communication Quarterly*, 84(1), 90–104.
- Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, 2(1), 13–43.
- Reinard, J. C. (2006). *Communication research statistics*. Sage Publications, Thousand Oaks, CA.
- Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, 12(3), 287–297.
- Reynolds, N., Diamantopoulos, A., & Schlegelmilch, B. (1993). Pretesting in questionnaire design: A review of the literature and suggestions for further research. *Journal of Market Research Society*, 35(2), 171–182.
- Riffe, D., Lacy, S., & Fico, F. G. (2005). *Analyzing media messages. Using quantitative content analysis in research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Riffe, D., Lacy, S., & Fico, F. G. (2014). *Analyzing media messages. Using quantitative content analysis in research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Ruel, E. E., Wagner, W. E., III, & Gillespie, B. J. (2016). *The practice of survey research*. Thousand Oaks, CA: Sage.
- Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor Factor analysis and scale revision. *Psychological Assessment*, 12(3), 1629–1646.
- Schreiber, J. B., Stage, F. K., King, J., Nora, A., & Barlow, E. A. (2006). Reporting structural equation modeling and confirmatory analysis results: A review. *The Journal of Educational Research*, 99(6), 323–337.
- Simms, L. J. (2008). Classical and modern of psychological scale construction. *Social and Psychology Compass*, 2(1), 414–433.
- Snook, S. C., & Gorsuch, R. L. (1989). Common factor analysis vs. component analysis. *Psychological Bulletin*, 106, 148–154.
- Streiner, D. L. (1994). Figuring out factors: The use and misuse of factor analysis. *Canadian Journal of Psychiatry*, 39, 135–140.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics* (5th ed.). New York: Allyn and Bacon.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn and Bacon.
- Thompson, B. (2004). *Exploratory and confirmatory analysis: Understanding concepts and applications*. Washington, DC, USA: American Psychological Association.
- Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL, USA: University of Chicago Press.
- Tinsley, H. E. A., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology*, 34(4), 414–424.

- Trumbo, C. W. (2004). Research methods in mass communication research: A census of eight journals 1990-2000. *Journalism & Mass Communication Quarterly*, 81(2), 417-436.
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3), 321-327.
- Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin, & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas Jackson at seventy* (pp. 41-71). Boston, MA: Kluwer.
- Viswanathan, M. (2010). Understanding the intangibles of measurement in the social sciences. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The SAGE handbook of measurement* (pp. 285-313). London, UK: Sage.
- Watkins, M. W. (2006). Determining parallel analysis criteria. *Journal of Modern Applied Statistical Methods*, 5(2), 344-346.
- Wimmer, R. D., & Haynes, R. B. (1978). Statistical analyses in the *Journal of Broadcasting*, 1970-76. *Journal of Broadcasting*, 22(2), 241-248.
- Worthington, R. L., & Whittaker, T. A. (2006). Scale development research. A content analysis for recommendations for best practices. *The Counseling Psychologist*, 34(6), 806-838.
- Ye, L., & Ki, E. (2012). The status of online public relations research: An analysis of published articles in 1992-2009. *Journal of Public Relations Research*, 24(5), 409-434.
- Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules determining the number of components to retain. *Multivariate Behavioral Research*, 17, 253-269.

REVIEW

Open Access



Scale development: ten main limitations and recommendations to improve future research practices

Fabiane F. R. Morgado^{1*}, Juliana F. F. Meireles², Clara M. Neves², Ana C. S. Amaral³ and Maria E. C. Ferreira²

Abstract

The scale development process is critical to building knowledge in human and social sciences. The present paper aimed (a) to provide a systematic review of the published literature regarding current practices of the scale development process, (b) to assess the main limitations reported by the authors in these processes, and (c) to provide a set of recommendations for best practices in future scale development research. Papers were selected in September 2015, with the search terms “scale development” and “limitations” from three databases: Scopus, PsycINFO, and Web of Science, with no time restriction. We evaluated 105 studies published between 1976 and 2015. The analysis considered the three basic steps in scale development: item generation, theoretical analysis, and psychometric analysis. The study identified ten main types of limitation in these practices reported in the literature: sample characteristic limitations, methodological limitations, psychometric limitations, qualitative research limitations, missing data, social desirability bias, item limitations, brevity of the scale, difficulty controlling all variables, and lack of manual instructions. Considering these results, various studies analyzed in this review clearly identified methodological weaknesses in the scale development process (e.g., smaller sample sizes in psychometric analysis), but only a few researchers recognized and recorded these limitations. We hope that a systematic knowledge of the difficulties usually reported in scale development will help future researchers to recognize their own limitations and especially to make the most appropriate choices among different conceptions and methodological strategies.

Keywords: Assessment, Measurement, Psychometrics, Reliability, Validity

Introduction

In recent years, numerous measurement scales have been developed to assess attitudes, techniques, and interventions in a variety of scientific applications (Meneses et al. 2014). Measurement is a fundamental activity of science, since it enables researchers to acquire knowledge about people, objects, events, and processes. Measurement scales are useful tools to attribute scores in some numerical dimension to phenomena that cannot be measured directly. They consist of sets of items revealing levels of theoretical variables otherwise unobservable by direct means (DeVellis 2003).

A variety of authors (Clark and Watson 1995; DeVellis 2003; Nunnally 1967; Pasquali 2010) have agreed that the scale development process involves complex and systematic procedures that require theoretical and methodological rigor. According to these authors, the scale development process can be carried out in three basic steps.

In the first step, commonly referred as “item generation,” the researcher provides theoretical support for the initial item pool (Hutz et al. 2015). Methods for the initial item generation can be classified as deductive, inductive, or a combination of the two. *Deductive* methods involve item generation based on an extensive literature review and pre-existing scales (Hinkin 1995). On the other hand, *inductive* methods base item development on qualitative information regarding a construct obtained from opinions gathered from the target population—e.g., focus groups,

* Correspondence: fabi.frm@hotmail.com

¹Institute of Education, Universidade Federal Rural do Rio de Janeiro, BR-465, km 7, Seropédica, Rio de Janeiro 23890-000, Brazil

Full list of author information is available at the end of the article

interviews, expert panels, and qualitative exploratory research methodologies (Kapuscinski and Masters 2010). The researcher is also concerned with a variety of parameters that regulate the setting of each item and of the scale as a whole. For example, suitable scale instructions, an appropriate number of items, adequate display format, appropriate item redaction (all items should be simple, clear, specific, ensure the variability of response, remain unbiased, etc.), among other parameters (DeVellis 2003; Pasquali 2010).

In the second step, usually referred to as the “theoretical analysis,” the researcher assesses the content validity of the new scale, ensuring that the initial item pool reflects the desired construct (Arias et al. 2014). A content validity assessment is required, since inferences are made based on the final scale items. The item content must be deemed valid to instill confidence in all consequent inferences. In order to ensure the content validity, the researcher seeks other opinions about the operationalized items. The opinions can be those of expert judges (experts in the development scales or experts in the target construct) or target population judges (potential users of the scale), enabling the researcher to ensure that the hypothesis elaborated in the research appropriately represents the construct of interest (Nunnally 1967).

In the last step, psychometric analysis, the researcher should assess whether the new scale has construct validity and reliability. Construct validity is most directly related to the question of what the instrument is in fact measuring—what construct, trait, or concept underlies an individual’s performance or score on a measure (Churchill 1979). This refers to the degree to which inferences can be legitimately made from the observed scores to the theoretical constructs about which these observations are supposed to contain information (Podsakoff et al. 2013). Construct validity can be assessed with the use of exploratory factor analysis (EFA), confirmatory factor analysis (CFA), or with convergent, discriminant, predictive/nomological, criterion, internal, and external validity. In turn, reliability is a measure of score consistency, usually measured by use of internal consistency, test-retest reliability, split-half, item-total correlation/inter-item reliability, and inter-observer reliability (DeVellis 2003). To ensure construct validity and reliability, the data should be collected in a large and appropriately representative sample of the target population. It is a common rule of thumb that there should be at least 10 participants for each item of the scale, making an ideal of 15:1 or 20:1 (Clark and Watson 1995; DeVellis 2003; Hair Junior et al. 2009).

Although the literature on theoretical and methodological care in scale development is extensive, many limitations have been identified in the process. These include failure to adequately define the construct domain,

failure to correctly specify the measurement model, underutilization of some techniques that are helpful in establishing construct validity (MacKenzie et al. 2011), relatively weak psychometric properties, applicability to only a single form of treatment or manual, extensive time required to fill out the questionnaire (Hilsenroth et al. 2005), inappropriate item redaction, too few items and participants in the construction and analysis, an imbalance between items that assess positive beliefs and those that assess negative beliefs (Prados 2007), social desirability bias (King and Bruner 2000), among others.

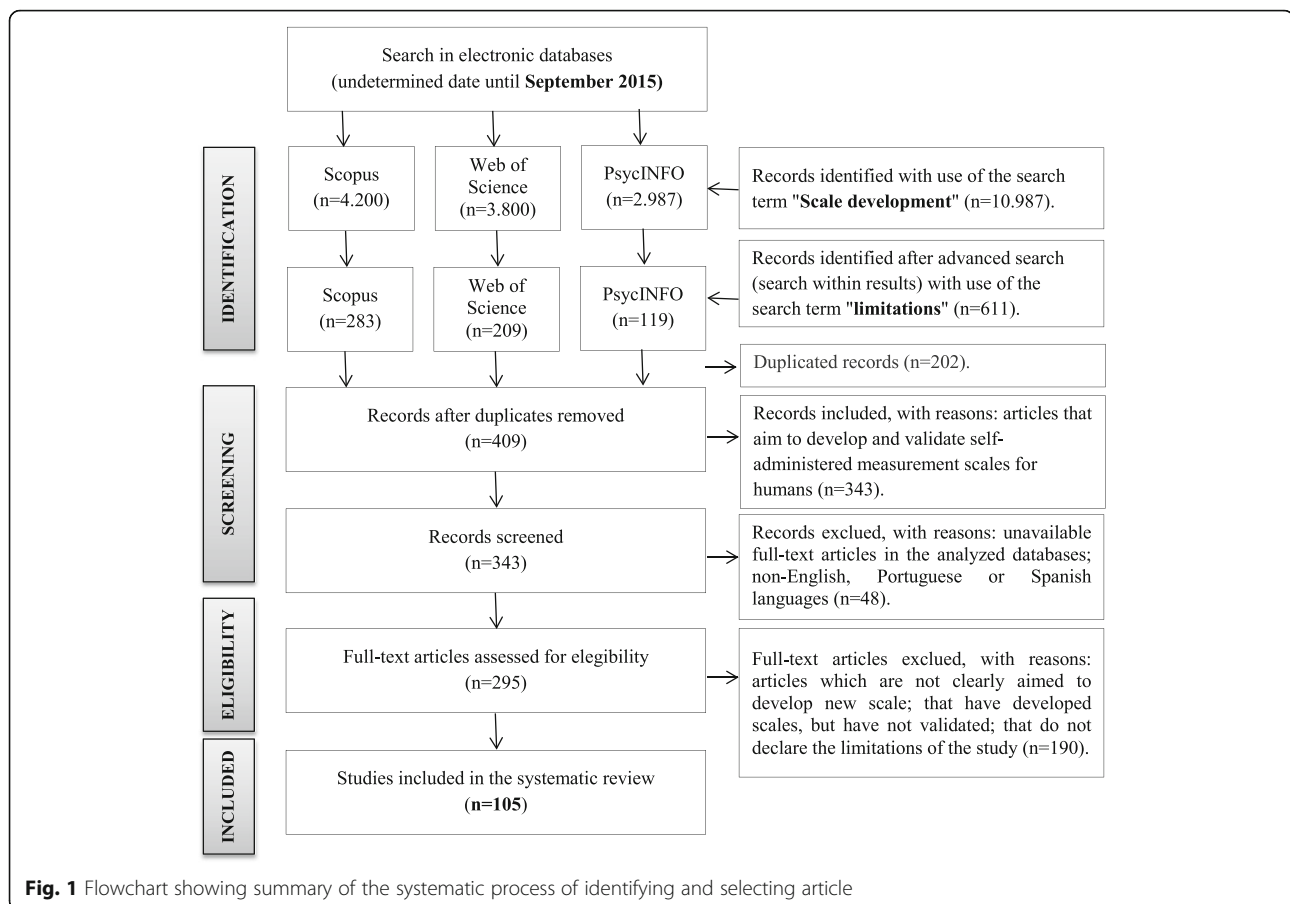
These limitations in the scale development process weaken the obtained psychometric results, limiting the future applicability of the new scale and hindering its generalizability. In this sense, knowledge of the most often reported limitations is fundamental in providing essential information to help develop best practices for future research in this area. The purpose of this article is threefold: (a) to provide a systematic review of the published literature regarding some current practices of the scale development process, (b) to assess the main limitations reported by the authors in this process, and (c) to provide a set of recommendations for best practices in future scale development research.

Review

Method

This systematic review identified and selected papers from three databases: Scopus, PsycINFO, and Web of Science. There was no time restriction in the literature search, which was completed in September 1, 2015. The following search term was used: “scale development.” In the set of databases analyzed, the search was done inclusively in “Any Field” (PsycINFO), in “Article Title, Abstract, Keywords” (Scopus), or in any “Topic” (Web of Science). In addition, we used an advanced search to filter the articles in (search within results), with the search term “limitations” identified in “Any Field” in all databases. Both terms were used in English only. Four reviewers evaluated the papers in an independent and blinded way. Any disagreements on eligibility of a particular study were resolved through consensus among reviewers.

Figure 1 shows a flowchart summarizing the strategy adopted for identification and selection of studies. We used only one inclusion criteria for the evaluation of the studies: (a) articles that aim to develop and validate self-administered measurement scales for humans. We excluded (a) unavailable full-text papers in the analyzed databases, (b) papers in languages other than English, Portuguese, or Spanish, (c) articles which were not clearly aimed at the development of a new scale (i.e., we excluded articles investigating only the reliability, validity, or revisions of existing scales and studies that describe



the validation of instruments for other languages), (d) papers with unvalidated scales, and (e) articles that did not declare the limitations of the study.

Results

In all, this systematic review evaluated 105 studies published between 1976 and 2015. Most (88.5%) was published between 2005 and 2015, and only two studies date from the last century. We analyzed two major issues: (a) *current practices of the scale development process*—considering the three steps usually reported in the literature (step 1—item generation, step 2—theoretical analysis, step 3—psychometric analysis), the number of participants in step 3, the number of items in the beginning scale, and the number of items in the final scale; (b) *main limitations reported by the authors in the scale development process*—considering the limitations observed and recorded by the authors during the scale development process. The description of these results can be found in Table 1.

Current practices of the scale development process

Step 1—item generation In the first step, 35.2% ($n = 37$) of the studies reported using exclusively deductive methods

to write items, 7.6% ($n = 8$) used only inductive methods, and 56.2% ($n = 59$) combined deductive and inductive strategies. The majority of the studies used a literature review (84.7%, $n = 89$) as the deductive method in item generation. In inductive methods, 26.6% of studies ($n = 28$) chose to conduct an interview.

Step 2—theoretical analysis In order to theoretically refine the items, several studies used opinions of experts (74.2%, $n = 78$), whereas others used target population opinions (43.8%, $n = 46$). In addition, 63.8% ($n = 67$) of the studies used only one of these approaches (expert or population judges).

Step 3—psychometric analysis The most common analyses that have been used to assess construct validity are EFA (88.6%, $n = 93$), CFA (72.3%, $n = 76$), convergent validity (72.3%, $n = 76$), and discriminant validity (56.2%, $n = 59$). Most studies opted to combine EFA and CFA (65.7%, $n = 69$). Only 4.7% ($n = 5$) failed to use factor analysis in their research. In relation to study reliability, internal consistency checks were used by all studies and test-retest reliability was the second most commonly used technique (22.8%, $n = 24$).

Table 1 Systematic review of the scale development process recorded in 105 included studies

Study	Scale	Step 1	Step 2	Step 3	N step 3	Initial item pool	Final item pool	Main limitations reported
Aagja and Garg (2010)	PubHosQual Scale	LR/ES/I	EJ	EFA/CFA/NV/CV/DV/ICR	401	59	24	LG
Ahmad et al. (2009)	Service Quality Scale	LR/FC	EJ	CFA/CV/DV/S-RR/ICR	413	31	10	LG
Akter et al. (2013)	MHealth Service Quality Scale	LR/ES/FC/I	EJ	EFA/CFA/NV/PV/CV/DV/I-JR/I-T-CR/ICR	305	29	22	LG/CSM
Alvarado-Herrera et al. (2015)	CSRConsPerScale	LR/ES	EJ	CFA/CV/DV/NV/ICR	1087	73	18	LG/Lack of the PV
Armfield (2010)	IDAF-4C ⁺	LR/ES	EJ	EFA/CV/PV/T-RR/ICR	1083	29	8	LG/Lack of the CV/SRM
Atkins and Kim (2012)	Smart Shopping Scale	LR/FC/I	EJ/TPJ	EFA/CFA/NV/CV/DV/ICR	1,474	62	15	LG
Bagdare and Jain (2013)	Retail Customer Experience Scale	LR/EP	EJ/TPJ	EFA/CFA/CV/ICR	676	45	12	LG/This study has not established DV and NV
Bakar and Mustafa (2013)	Organizational Communication Scale	LR/FC	EJ	EFA/CFA/CV/CV/DV/T-RR/ICR	596	386	38	LG/Inadequate choose variables to be correlated
Beaudreuil et al. (2011)	URAM	EP/I	EJ/TPJ	EFA/CV/DV/T-RR/ICR	138	52	9	LG/SSS
Bhattacharjee (2002)	Individual trust in online firms scale	LR/ES	TPJ	CFA/CV/DV/NV/ICR	269	18	7	WBS
Blankson et al. (2007)	International consumers selection of banks scale	LR/FC	EJ/TPJ	EFA/CFA/CV/PV/NV/ICR/I-T-CR	773	60	18	LG
Blankson et al. (2012)	Scale measuring college students' choice criteria of credit cards	FC/ES	EJ	EFA/CFA/CV/DV/S-RR/ICR	405	59	19	LG/CSM
Bolton and Lane (2012)	IEO	LR/ES	TPJ	EFA/NV/EV/CV/DV/I-T-CR/ICR	1162	NCR	10	LG/Lack of the CFA
Bova et al. (2006)	HCR	I/FC	EJ	EFA/T-RR/ICR	99	58	15	LG/Scale was administered in a face-to-face interview/SSS.
Boyar et al. (2014)	CESS	LR	EJ	CFA/DV/ICR	446	140	52	CSM
Brock and Zhou (2005)	OIU	LR/I	EJ	DV/PV/NV/ICR	112	NCR	7	LG
Brun et al. (2014)	Online relationship quality scale	LR, and ES	EJ/TPJ	EFA/CFA/CV/DV/PV/ICR	476	33	21	LG
Butt and Run (2010)	SERVQUAL model scale	LR/EP	EJ	EFA/CFA/CV/DV/ICR	340	17	17	LG
Caro and García (2007)	Perceived service quality in urgent transport service scale	LR/ES	EJ/TPJ	EFA/CFA/DV/CV/NV/I-T-CR/ICR	375	68	38	LG/Lack of the CV or DV
Chahal and Kumari (2012)	CPV Scale	LR/ES	EJ/TPJ	EFA/CFA/CV/I-T-CR/ICR	515	32	27	LG
Chen et al. (2009)	Process Orientation Scale	LR/I	EJ	EFA/CFA/CV/DV/I-I-CR/ICR	356	NCR	6	LG/SSS/Lack of the NV
Choi et al. (2011)	Measure of dyspnea severity and related functional limitations	LR/EP	EJ/TPJ	EFA/CFA/CV/DV/T-RR/ICR	608	364	33	CSM
		LR/EP/ES	EJ	CFA/CV/PV/EV/ICR	378	80	54	CSM/SRM

Table 1 Systematic review of the scale development process recorded in 105 included studies (Continued)

Christophersen and Konradt (2012)	Reflective and formative usability scales	LR/EP	EJ/TPJ	EFA/CFA/CV/DV/ICR	1281	NCR	29	Items are not reverse-scored
Cicero et al. (2010)	ASI	LR/EP	EJ/TPJ	EFA/CFA/CV/DV/ICR	1200	65	11	LG
Coker et al. (2011)	IPPR	LR/I	EJ/TPJ	EFA/CFA/CV/NV/DV/ICR	210	119	15	LG/Deductive approach to scale development
Coleman et al. (2011)	B2B service brand identity scale	LR	EJ/TPJ	EFA/CFA/DV/I-T-CR/ICR	201	30	17	LG/CSM
Colwell et al. (2008)	Measure of service convenience	LR/I	EJ/TPJ	EFA/CFA/CV/DV/NV/ICR	332	121	70	CSM
Cossette et al. (2005)	Caring Nurse–Patient Interactions Scale	LR/ES	EJ	EFA/CV/CtV/ICR	293	71	42	MD
Dennis and Bocarnea (2005)	Servant leadership assessment instrument	LR/EP	EJ	EFA/CtV/ICR	3100	9	8	LG
Devlin et al. (2014)	Fairness measurement scale	LR, and ES	EJ/TPJ	EFA/CFA/CV/DV/NV/ICR	431	72	21	SRM
Dunham and Burt (2014)	Organizational Memory Scale	LR/ES	NCR	EFA/CFA/CV/T-RR/ICR	270	NCR	84	LG
Edwards et al. (2010)	STL	FC	TPJ	EFA/CV/CtV/ICR	282	136	136	LG/SSS
Feuerstein et al. (2005)	Response to work in those with upper extremity pain scale	LR/FC/ES	TPJ	EFA/T-RR/ICR	213	9	4	SSS/Subjective Analysis/SRM
Fisher et al. (2014)	Entrepreneurial Success Scale	LR/I	EJ	EFA/CFA/ICR	430	122	43	LG
Flight et al. (2011)	Characteristics-based innovation adoption scale	LR	EJ/TPJ	EFA/CFA/ICR/EP/CV/DV/NV/ICR	1528	160	45	LG
Forbush et al. (2013)	EPSI	LR	NCR	EFA/CFA/CV/CtV/DV/T-RR/ICR	2259	35	33	Lack of the validity
Foster et al. (2015)	GNS	LR/ES	NCR	EFA/CFA/ICR	632	NCR	22	SSS/CSM
Franchette et al. (2007)	RRTW	LR/EP	EJ	EFA/CFA/CV/PV/IV/EP/ICR	592	79	54	LG
Gesten (1976)	HRI	LR/EP/ES	EJ	EFA/T-RR/ICR	490	53	22	LG
Gibbons et al. (2013)	MULTIPLEs	LR/ES/QR	TPJ	EFA/T-RR/ICR	151	NCR	21	CSM
Gilgor and Holcomb (2014)	SCA	LR/ES/I	EJ/TPJ	EFA/CFA/CV/DV/EP/ICR	1496	26	10	LG/MD
Glynn et al. (2015)	PFS	ES/QR	NCR	EFA/CV/T-RR/ICR	739	13	11	LG/Items ambiguous/Difficult to control variables
Gottlieb et al. (2014)	Consumer perceptions of trade show effectiveness scale	LR/I	NCR	EFA/CFA/CV/DV/NV/I-T-CR/ICR	502	25	11	LG/CSM
Hall et al. (2002)	General Trust in physicians scale	LR/FC/EP	EJ/TPJ	EFA/CV/CtV/ICR	401	NCR	17	LG
Han et al. (2011)	Scale of switching barriers in full-service restaurants	LR/FC	EJ/TPJ	EFA/CFA/CV/NV/I-JR/ICR	1288	26	15	LG
Henderson-King and Henderson-King (2005)	ACSS	LR	TPJ	EFA/DV/CV/T-RR/ICR				

Table 1 Systematic review of the scale development process recorded in 105 included studies (Continued)

Hernandez and Santos (2010)	Development-based trust	LR/I	TPJ	EFA/CFA/CV/DV/NV/ICR	238	30	27	CSM	
Hildebrandt et al. (2004)	MDDI	LR/ES	NCR	EFA/CV/DV/T-RR/ICR	245	21	20	LG/Lack of the DV	
Ho and Lin (2010)	Scale for measuring internet banking service quality	LR/IES	TPJ	EFA/DV/CV/ICR	130	30	17	SSS	
Jong et al. (2014)	CRIQ	LR/ES	EJ	EFA/CFA/T-RR/ICR	310	120	120	Lack of the CFA - the CFI fit is below the 0.90	
Kim et al. (2011)	CEI	LR	TPJ	EFA/CFA/CV/DV/ICR	397	134	26	LG/Lack of the validity/WBS	
Kim et al. (2014)	SAPS	LR	EJ	CFA/CV/CV/ICR	795	29	15	Lack of the DV	
Kwon and Lennon (2011)	Brand Association Scale	LR	EJ	EFA/CFA/CV/DV/I-IR/ICR	671	28	14	LG	
Lin and Hsieh (2011)	SSTQUAL Scale	LR/FC/I	EJ	EFA/CFA/CV/DV/NV/I-T-CR/ICR	1238	75	20	LG/subjectivity in the EFA and CFA	
Lombaerts et al. (2009)	SRLTB	LR	EJ	EFA/CFA/ICR	952	39	10	Initial unsatisfactory factor analysis output	
Lucas-Carrasco et al. (2011)	QOCS	LR/FC	TPJ	EFA/CFA/CV/DV/ICR	3772	44	17	Recruitment of a larger number of interviewers	
Mahudin et al. (2012)	Measuring rail passenger crowding	LR/ES	EJ/TPJ	EFA/CFA/CV/DV/ICR	525	9	20	Lack of the CTV/SRM	
Medina-Pradas et al. (2011)	BDSEE	ES/EP	EJ	EFA/CV/CV/ICR	77	14	14	SSS/CSM	
Morean et al. (2012)	AEAS	LR/ES	EJ/TPJ	EFA/CFA/CV/CV/DV/T-RR/ICR	546	40	22	LG/SRM/CSM	
Morgado et al. (2014)	SAS-EB	LR/FC	EJ/TPJ	CFA/CV/DV/ICR	318	33	18	Lack of the use of a validated scale in the CV	
Nagy et al. (2014)	Scale to measure liabilities and assets of newness after start-up	LR/I	EJ	EFA/CFA/DV/ICR	260	235	19	LG/SSS	
Napoli et al. (2014)	Consumer-based brand authenticity scale	LR	EJ/TPJ	EFA/CFA/CV/DV/PV/ICR	762	157	14	Lack of a more robust LR	
Negra and Mzoughi (2012)	OCPS	LR/I	EJ	EFA/CFA/CV/DV/NV/I-T-CR/ICR	512	77	5	Widely heterogeneous sample/Brevity of the scale.	
Ngorsuraches et al. (2007)	TRUST-Ph	LR/FC/EP/ES	EJ	EFA/CV/CV/ICR	400	40	30	LG/SSS/MD/social desirability bias	
Oh (2005)	Affective reactions to print apparel advertisements scale	LR/FC/ES	TPJ	EFA/CFA/CV/DV/CV/ICR	128	66	54	LG	
Olaya et al. (2012)	ISAD	EP/ES	EJ	CV/DV/T-RR/ICR	76	20	17	LG	
Omar and Musa (2011)	LPSQual	LR/FC	EJ	EFA/CFA/CV/DV/NV/ICR	655	57	26	LG/Lack of the NV/CSM	
Pan et al. (2013)	PMGS	I	EJ/TPJ	EFA/CFA/CV/S-RR/I-T-CR/I-T-CR/ICR	554	71	14	LG/SRM/Lack of the T-RR	
Patwardhan and Balasubramanian (2011)	Measurement scale for brand romance	LR/ES/QER	TPJ	EFA/CFA/CV/DV/CV/NV/ICR	711	70	12	LG	
Pimentel et al. (2007)	EPM	NCR	EJ/TPJ	EFA/CFA/ICR	480	13	13	LG/Lack of the CV and DV	

Table 1 Systematic review of the scale development process recorded in 105 included studies (Continued)

	PCQ	EP/FC	TPJ	EFA/CFA/CV/ICR	953	391	18	CSM
Pommer et al. (2013)	ESLS	ES	EJ	EFA/CFA/CV/DV/ICR	218	55	25	LG/SRM/WBS
Reed et al. (2011)	MDRS-22	LR	EJ	EFA/CFA/ICR	1176	82	22	LG/Lack of the T-RR/Lack of the CV
Rice et al. (2013)	RSM-scale	LR/ES	EJ/TPJ	DV/T-RR/ICR	136	43	36	LG
Rodrigues and Bastos (2012)	Organizational Entrenchment Scale	EP/ES	EJ	EFA/CFA/I-T-CR/I-H-CR/ICR	721	31	22	LG
Rodríguez et al. (2013)	VEDAS	ES	NCR	EFA/CFA/CV/CV/T-RR/ICR	1034	40	20	Long time between the test and retest/Lower Cronbach's alpha
Rosenthal (2011)	IEK	EP	EJ	EFA/CV/CV/I-T-CR/I-H-T-RR/ICR	292	54	21	LG/SSS
Saxena et al. (2015)	UCLA Hoarding Severity Scale	LR/EP	EJ	EFA/CV/DV/I-H-CR/ICR	127	NCR	10	Lack of the T-RR/Lack of the instructions for raters in the initial version of the scale
Schlosser and McNaughton (2009) -	I-MARKOR scale	LR/FC/I	EJ/TPJ	EFA/CFA/CV/DV/NV/ICR	138	71	20	SSS/CSM.
Sewitch et al. (2003)	PPDS	LR	EJ	EFA/CV/ICR	200	10	10	LG/CrV was limited/content validity was not formally assessed
Sharma (2010)	Personal Cultural Orientations Scale	LR/I	EJ	EFA/CFA/NV/CV/PV/DV/ICR	2332	96	40	LG/Lack of the PV
Sharma and Gassenheimer (2009)	SPC Scale	LR/EP/I	EJ	EFA/CFA/CV/DV/ICR	511	8	17	Lack of the EV
Shawyer et al. (2007)	VAAS	LR	EJ	CV/T-RR/ICR	41	61	31	Lack of a more robust demonstration of the validity/SSS
Sin et al. (2005)	CRM	LR	EJ	EFA/CFA/CV/DV/NV/ICR	641	78	18	LG/CSM
Sohn and Choi (2014)	Expected Interactivity Scale	LR/EP/I	EJ	EFA/CFA/CV/DV/CV/T-RR/ICR	378	50	12	Lack of the empirical test
Song et al. (2011)	SECI	LR	EJ	EFA/CFA/CV/I-H-CR/ICR	469	26	17	LG/deductive approach
Staines (2013)	Investigative Thinking Styles Scale	LR	TPJ	EFA/CV/CV/ICR	545	68	16	LG
Sultan and Wong (2010)	Performance-based service quality model scale	LR/FC/ES	EJ	EFA/CFA/ICR	362	67	67	The study uses three sources to collect data
Swaid and Wigand (2009)	E-Service Quality Scale	LR	TPJ	EFA/CFA/CV/DV/ICR.	557	NCR	28	Online survey
Tanimura et al. (2011)	DDLKOS	LR/I	EJ	EFA/CFA/CV/ICR	362	48	14	Inadequate choose variables to be correlated with that of the study
Taute and Sierra (2014)	Brand Tribalism Scale	LR/ES	NCR	EFA/CFA/CV/DV/ICR	616	35	16	LG
Tombaugh et al. (2011)	SEW	LR/EP	NCR	EFA/CFA/CV/DV/PV/ICR	348	5	5	CSM/Brevity of the scale
Turker (2009)	CSR	LR/FC/ES	TPJ	EFA/I-H-CR/I-T-CR/ICR	269	55	18	LG
Uzunboyulu and Ozdamli (2011)	MLPS	LR/EP	EJ	EFA/S-RR/ICR	467	31	26	LG

Table 1 Systematic review of the scale development process recorded in 105 included studies (Continued)

Van der Gaag et al. (2013)	DACOB5	EP	EJ	EFA/CV/ICR/S-RR/T-RR	257	70	42	SSS/Validation performed with patients/inappropriate choice of the instruments for validation
Von Steinbüchel et al. (2010)	QOLIBRI	LR/ES	EJ	EFA/CFA/T-RR/ICR	2449	148	37	SSS
Voon et al. (2014)	HospISE	LR/FC	EJ/TPJ	EFA/CFA/CV/DV/CV/ICR	1558	NCR	21	LG/CSM
Walshe et al. (2009)	DIP	LR/ES	TPJ	Ecological validity/ICR	31	48	48	SSS/Lack of the DV, CV and T-RR
Wang and Mowen (1997)	SC	LR	EJ	EFA/CFA/CV/DV/PV/I-T-CR/ICR	140	60	9	SSS
Wepener and Boshoff (2015)	The customer-based corporate reputation of large service organizations scale	LR/ES/FC	EJ	EFA/CFA/NV/CV/DV/ICR	2551	78	19	LG
Williams et al. (2009)	SCSC	LR/I	EJ/TPJ	EFA/CFA/CV/DV/PV/I-T-CR/ICR	162	5	5	LG; b) WBS.
Wilson and Holmvall (2013)	Incivility from customers scale	LR/FC	EJ	EFA/CFA/CV/DV/CV/ICR	439	27	10	LG/CSM/SRM
Yang et al. (2014)	BLOG-S-INNO Scale	EP	TPJ	EFA/CFA/CV/DV/ICR	498	517	18	LG
Zhang and Hu (2011)	Farmer-buyer relationships in China Scale	LR/ES	EJ/TPJ	EFA/CFA/CV/I-I-CR/ICR	210	39	22	LG
Zheng et al. (2010)	DPEBBS	LR/FC	EJ	EFA/CFA/CV/T-RR/I-T-CR/ICR	269	51	24	LG/SSS/EFA and CFA - same sample/Reliability coefficients - unsatisfactory.

N sample size, *EFA* exploratory factor analysis, *CFA* confirmatory factor analysis, *NV* nomological validity, *CV* convergent validity, *CrV* concurrent validity, *CtV* criterion validity, *DV* discriminant validity, *DIV* divergent validity, *PV* predictive validity, *IV* internal validity, *EV* external validity, *ICR* internal consistency reliability, *S-RR* split-half reliability, *I-JR* inter-judge reliability, *I-T-CR* item-total correlation reliability, *I-I-CR* inter-item correlation reliability, *T-RR* test-retest reliability, *LR* literature review, *ES* existing scales, *I* interview, *FC* Focus group, *EP* expert panel, *QER* qualitative exploratory research, *NCR* not clearly reported, *EJ* expert judges, *TPJ* target population judges, *LG* limitations of generalization, *SSS* small sample size, *CSM* cross-sectional methodology, *SEM* self-reporting methodology, *WBS* web-based survey, *MD* Missing data

Sample size in step 3 and number of items Interestingly, 50.4% ($n = 53$) of the studies used sample sizes smaller than the rule of thumb, which is a minimum of 10 participants for each item in the scale. Regarding number of items, the majority of the studies (49.6%, $n = 52$) lost more than 50% of the initial item pool during the validation process.

Table 2 summarizes and provides more details on our findings regarding the current practices in the scale development.

Main limitations reported in the scale development process

As result of this systematic review, we found ten main limitations commonly referenced in the scale development process: (1) sample characteristic limitations—cited by 81% of the studies, (2) methodological limitations—33.2%, (3) psychometric limitations—30.4%, (4) qualitative research limitations—5.6%, (5) missing data—2.8%, (6) social desirability bias—1.9%, (7) item limitations—1.9%, (8) brevity of the scale—1.9%, (9) difficulty controlling all variables—0.9%, and (10) lack of manual instructions—0.9%. Table 3 summarizes these findings.

Discussion

This systematic review was primarily directed at identifying the published literature regarding current practices of the scale development. The results show a variety of practices that have been used to generate and assess items, both theoretically and psychometrically. We evaluated these current practices, considering three distinct steps (item generation, theoretical analysis, and psychometric analysis). We also considered the relationship between sample size and number of items, since this is considered an important methodological aspect to be evaluated during the scale development process. The results are discussed together with recommendations for best practices in future scale development research.

Current practices of the scale development process—findings and research implications

Regarding step 1, item generation, our results show that, although several studies used exclusively deductive methods (e.g., Henderson-King and Henderson-King 2005; Kim et al. 2011), the majority (e.g., Bakar and Mustaffa 2013; Uzunboyu and Ozdamli 2011) combined deductive and inductive methods, a combination consistent with the recommended strategy for the creation of new measures (DeVellis 2003). These findings, however, differ from previous critical reviews of scale development practices, which found that most of the reported studies used exclusively deductive methods (Hinkin 1995; Kapuscinski and Masters 2010; Ladhari 2010). This is particularly important since the quality of

Table 2 Summary of current practices of the scale development process

Methods	Number of scales resorting to	Percentage (%) of scales resorting to
Step 1—item generation		
Deductive methods (exclusively)	37	35.2
Inductive methods (exclusively)	8	7.6
Combined deductive and inductive methods	59	56.2
Literature review	89	84.7
Existing scales	40	38
Interviews	28	26.6
Focus groups	25	23.8
Expert panel	23	21.9
Qualitative exploratory research	3	5
Not clearly reported method	1	1
Step 2—theoretical analysis		
Expert judges	78	74.2
Target population judges	46	43.8
Use of just one approach	67	63.8
Combined two approaches	29	27.7
Not clearly reported approach	9	8.5
Step 3—psychometric analysis		
EFA	93	88.6
CFA	76	72.3
Combined EFA and CFA	69	65.7
Lack of EFA and CFA	5	4.7
Convergent/concurrent validity	76	72.3
Discriminant validity	59	56.2
Predictive/nomological validity	34	32.3
Criterion validity	17	16.2
External validity	5	4.7
Internal validity	3	2.8
Internal consistency	105	100
Test-retest reliability	24	22.8
Item-total correlation/inter-item reliability	19	18.1
Split-half reliability	3	2.9
Inter-judge reliability	3	2.9
Sample size about step 3 and number of items		
Sample size smaller than the rule of thumb 10:1	53	50.4
Number of items final scale reduced by 50%	42	40
Number of items final scale reduced more than 50%	52	49.6
Not clearly reported initial item number	11	10.4

EFA exploratory factor analysis, CFA confirmatory factor analysis

Table 3 Scale development process—ten main limitations

	Limitations	<i>n</i>	%
1	Sample characteristics limitations	85	81
	Homogeneous and/or convenience sample—limitations of generalization	67	64
	Small sample size	18	17
2	Methodological limitations	35	33.2
	Cross-sectional methodology	20	19
	Self-reporting methodology	9	8.5
	Web-based survey	6	5.7
3	Psychometric limitations	32	30.4
	Lack of a more robust demonstration of the construct validity and/or reliability	21	20
	Inadequate choose of the instruments or variables to be correlated with the variable of the study	6	5.7
	Factor analysis limitations	5	4.7
4	Qualitative research limitations	6	5.6
	Deductive approach to scale development	2	1.9
	Lack of a more robust literature review	1	1
	Subjective analysis	1	0.9
	Content validity was not formally assessed	1	0.9
	Recruitment of a larger number of interviewers	1	0.9
5	Missing data	3	2.8
6	Social desirability bias	2	1.9
7	Items limitations	2	1.9
	Items ambiguous or difficult to answer	1	1
	None of the items are reverse-scored	1	0.9
8	Brevity of the scale	2	1.9
9	Difficult to control all variables	1	0.9
10	Lack of a manualized instructions	1	0.9

generated items depends on the way that the construct is defined. Failing to adequately define the conceptual domain of a construct causes several problems related to poor construct definition, leading to, for example, (a) confusion about what the construct does and does not refer to, including the similarities and differences between it and other constructs that already exist in the field, (b) indicators that may either be deficient or contaminated, and (c) invalid conclusions about relationships with other constructs (MacKenzie et al. 2011). Considering that item generation may be the most important part of the scale development process, future measures should be developed using the appropriate definition of the conceptual domain based on the combination of both deductive and inductive approaches.

Our results suggest that literature review was the most widely used deductive method (e.g., Bolton and Lane

2012; Henderson-King and Henderson-King 2005). This is consistent with the views of several other researchers who have systematically reviewed scales (Bastos et al. 2010; Ladhari 2010; Sveinbjornsdottir and Thorsteinsson 2008). Nevertheless, this finding differs from another study (Kapuscinski and Masters 2010) that found that the most common deductive strategies were reading works by spiritual leaders, theory written by psychologists, and discussion among authors. Literature review should be considered central for the enumeration of the constructs. It also serves to clarify the nature and variety of the target construct content. In addition, literature reviews help to identify existing measures that can be used as references to create new scales (Clark and Watson 1995; DeVellis 2003). In this sense, future research should consider the literature review as the initial and necessary deductive step foundational to building a new scale.

This review also highlights the fact that interviews and focus groups were the most widely used inductive methods (e.g., Lin and Hsieh 2011; Sharma 2010). Similar results were found in the systematic review by Kapuscinski and Masters (2010), Sveinbjornsdottir and Thorsteinsson (2008), and Ladhari (2010). These findings have particular relevance to future researchers, since they emphasize the importance of using methodological strategies that consider the opinions of the target population. Despite the fact that a panel of experts contributes widely to increasing the researchers' confidence in the content validity of the new scale, it is important to also consider the most original and genuine information about the construct of interest, which can be best obtained through reports obtained from interviews and focus groups with the target population.

Related to step 2, theoretical analysis, the results of this review indicate that expert judges have been the most widely utilized tool for analyzing content validity (e.g., Uzunboylu and Ozdamli 2011; Zheng et al. 2010). Previous studies have also found expert opinion to be the most common qualitative method for the elimination of unsuitable items (Kapuscinski and Masters 2010; Ladhari 2010). In the literature review conducted by Hardesty and Bearden (2004), the authors highlighted the importance of these experts to carefully analyze the initial item pool. They suggested that any research using new, changed, or previously unexamined scale items, should at a minimum be judged by a panel of experts. However, the authors also point out the apparent lack of consistency in the literature in terms of how researchers use the opinions of expert judges in aiding the decision of whether or not to retain items for a scale. Given this inconsistency, the authors developed guidelines regarding the application of different decision rules to use for item retention. For example, the "sumscore decision rule," defined as the total score for an item across all

judges, is considered by the authors to be the most effective in predicting whether an item should be included in a scale and appears, therefore, to be a reasonable rule for researchers to employ.

Future research in developing scales should be concerned, not only with opinions from experts but also with the opinions of the target population. The results of this review show that only a minority of studies considered the review of the scales' items by members of the target population (e.g., Uzunboylu and Ozdamli 2011; Zheng et al. 2010). In addition, a smaller minority combined the two approaches in the assessment of item content (e.g., Mahudin et al. 2012; Morgado et al. 2014). The limited use of target population opinions is a problem. A previous study of systematic scale development reviews found that the opinion of these people is the basis for content validity (Bastos et al. 2010). As highlighted by Clark and Watson (1995) and Malhotra (2004), it is essential for the new scale to undergo prior review by members of the target population. Pre-test or pilot study procedures make it possible to determine respondents' opinions of, and reactions to, each item on the scale, enabling researchers to identify and eliminate potential problems in the scale before it is applied at large.

Another problem noted in this systematic review was that some studies failed to clearly report how they performed the theoretical analysis of the items (e.g., Glynn et al. 2015; Gottlieb et al. 2014). We hypothesized that the authors either did not perform this analysis or found it unimportant to record. Future research should consider this analysis, as well as all subsequent analyses, necessary and relevant for reporting.

Almost all studies (95.3%) reported using at least one type of factor analysis—EFA or CFA—in step 3, psychometric analysis (e.g., Sewitch et al. 2003; Tanimura et al. 2011). Clark and Watson (1995) consider that “unfortunately, many test developers are hesitant to use factor analysis, either because it requires a relatively large number of respondents or because it involves several perplexing decisions” (p. 17). They emphasized the importance of the researcher's need to understand and apply this analysis, “it is important that test developers either learn about the technique or consult with a psychometrician during the scale development process” (Clark and Watson 1995, p. 17). This question seems to have been almost overcome in recent studies, since the vast majority of the analyzed studies used the factor analysis method.

Among the studies than used factor analysis, the majority chose to use EFA (e.g., Bakar and Mustaffa 2013; Turker 2009). Similar to our findings, Bastos et al. (2010) and Ladhari (2010) found EFA to be the more commonly utilized construct validity method when compared to CFA. EFA has extensive value because it is

considered to be effective in identifying the underlying latent variables or factors of a measure by exploring relationships among observed variables. However, it allows for more subjectivity in the decision-making process than many other statistical procedures, which can be considered a problem (Roberson et al. 2014).

For more consistent results on the psychometric indices of the new scale, DeVellis (2003) indicates the combined use of EFA and CFA, as was performed with most studies evaluated in this review. In CFA, the specific hypothesized factor structure proposed in EFA (including the correlations among the factors) is statistically evaluated. If the estimated model fits the data, then a researcher concludes that the factor structure replicates. If not, the modification indices are used to identify where constraints placed on the factor pattern are causing a misfit (Reise et al. 2000). Future studies should consider the combined use of EFA and CFA during the evaluation of construct validity of the new measure, and should also apply a combination of multiple fit indices (e.g., modification indices) in order to provide more consistent psychometric results.

After EFA and CFA, convergent validity was the preferred technique used in the vast majority of the studies included in this review (e.g., Brun et al. 2014; Cicero et al. 2010). This finding is consistent with prior research (Bastos et al. 2010). Convergent validity consists in examining whether a scale's score is associated with the other variables and measures of the same construct to which it should be related. It is verified either by calculating the average variance extracted for each factor when the shared variance accounted for 0.50 or more of the total variance or by correlating their scales with a measure of overall quality (Ladhari 2010). In the sequence of convergent validity, the following methods were identified as favorites in the assessment of construct validity: discriminant validity (the extent to which the scale's score does not correlate with unrelated constructs) (e.g., Coker et al. 2011), predictive/nomological validity (the extent to which the scores of one construct are empirically related to the scores of other conceptually related constructs) (e.g., Sharma 2010), criterion validity (the empirical association that the new scale has with a gold standard criterion concerned with the prediction of a certain behavior) (e.g., Tanimura et al. 2011), internal (signifies whether the study results and conclusions are valid for the study population), and external validity (generalizability of study) (e.g., Bolton and Lane 2012; Khorsan and Crawford 2014). Considering the importance of validity to ensure the quality of the collected data and the generalized potential of the new instrument, future studies should allow different ways to assess the validity of the new scale, thus increasing the psychometric rigor of the analysis.

With regard to reliability, all studies reported internal consistency statistics (Cronbach's alpha) for all subscales and/or the final version of the full scale (e.g., Schlosser and McNaughton 2009; Sewitch et al. 2003). These findings are consistent with those of previous review studies (Bastos et al. 2010; Kapuscinski and Masters 2010). DeVellis (2003) explains that internal consistency is the most widely used measure of reliability. It is concerned with the homogeneity of the items within a scale. Given its importance, future studies should to consider alpha evaluation as a central point of measurement reliability, and yet, as much as possible, involve the assessment of internal consistency with other measures of reliability. In the sequence of internal consistency, the following methods were identified by this review: test-retest reliability (analysis of the temporal stability; items are applied on two separate occasions, and the scores could be correlated) (e.g., Forbush et al. 2013), item-total/inter-item correlation reliability (analysis of the correlation of each item with the total score of the scale or subscales/analysis of the correlation of each item with another item) (e.g., Rodrigues and Bastos 2012), split-half reliability (the scale is split in half and the first half of the items are compared to the second half) (e.g., Uzunboylu and Ozdamli 2011), and inter-judge reliability (analysis of the consistency between two different observers when they assess the same measure in the same individual) (e.g., Akter et al. 2013; DeVellis 2003; Nunnally 1967).

Regarding sample size in step 3 and number of items, a particularly noteworthy finding was that most studies utilized sample sizes smaller than the rule of thumb that the minimum required ratio should be 10:1 (e.g., Turker 2009; Zheng et al. 2010). DeVellis (2003) and Hair Junior et al. (2009) comment that the sample size should be as large as possible to ensure factor stability. The 'observations to variables' ratio is ideal at 15:1, or even 20:1. However, most of the studies included in this review failed to adopt this rule. Some studies looked for justification on evidence related to the effectiveness of much smaller observations to variables ratios. For example, Nagy et al. (2014) justified the small sample size used in their investigation based on the findings of Barrett and Kline (1981), concluding that the difference in ratios 1.25:1 and 31:1 was not a significant contributor to results obtained in the factor stability. Additionally, Arrindell and van der Ende (1985) concluded that ratios of 1.3:1 and 19.8:1 did not impact the factor stability. Although the rules of thumb vary enormously, ten participants to each item has widely been considered safe recommended (Sveinbjornsdottir and Thorsteinsson 2008).

Finally, several studies had their number final of items reduced by more than 50%. For example, Flight et al. (2011) developed an initial item pool composed of 122

items and finished the scale with only 43. Pommer et al. (2013) developed 391 initial items and finished with only 18. Our findings clearly indicate that a significant amount of items can get lost during the development of a new scale. These results are consistent with previous literature which states both that the initial number of items must be twice the desired number in the final scale, since, during the process of analysis of the items, many may be excluded for inadequacy (Nunnally 1967), and that the initial set of items should be three or four times more numerous than the number of items desired, as a good way to ensure internal consistency of the scale (DeVellis 2003). Future research should consider these issues and expect significant loss of items during the scale development process.

Ten main limitations reported in the scale development process—findings and research implications

In addition to identifying the current practices of the scale development process, this review also aims to assess the main limitations reported by the authors. Ten limitations were found, which will be discussed together with recommendations for best practices in future scale development research (Table 3).

Sample characteristic limitations The above-mentioned limitations were recorded in the majority of the studies, in two main ways. The first and the most representative way was related to the sample type. Several studies used homogeneous sampling (e.g., Forbush et al. 2013; Morean et al. 2012), whereas others used convenience sampling (e.g., Coker et al. 2011; Flight et al. 2011). Both homogeneous and convenience samples were related to limitations of generalization. For example, Atkins and Kim (2012) pointed out that "the participants for all stages of the study were US consumers; therefore, this study cannot be generalized to other cultural contexts." Or indeed, "convenience samples are weaknesses of this study, as they pose generalizability questions," as highlighted by Blankson et al. (2012). Nunnally (1967) suggested that, to extend the generalizability of the new scale, sample diversification should be considered in terms of data collection, particularly in the psychometric evaluation step. Future studies should consider this suggestion, recruiting heterogeneous and truly random samples for the evaluation of construct validity and the reliability of the new measure.

The second way was related to small sample size. As previously described, most of the analyzed studies utilized sample sizes less than 10:1. Only some of the authors recognized this flaw. For example, Nagy et al. (2014) reported that "the sample size employed in conducting the exploratory factor analysis is another potential limitation of the study," Rosenthal (2011) described,

“the current study was limited by the relatively small nonprobability sample of university students,” and Ho and Lin (2010) recognized that “the respondent sample size was small.” Based in these results, we emphasize that future research should seek a larger sample size (minimum ratio of 10:1) to increase the credibility of the results and thus obtain a more exact outcome in the psychometric analysis.

Methodological limitations Cross-sectional methods were the main methodological limitations reported by other studies (e.g., Schlosser and McNaughton 2009; Tombaugh et al. 2011). Data collected under a cross-sectional study design contains the typical limitation associated with this type of research methodology, namely inability to determine the causal relationship. If cross-sectional methods are used to estimate models whose parameters do in fact vary over time, the resulting estimation may fail to yield statistically valid results, fail to identify the true model parameters, and produce inefficient estimates (Bowen and Wiersema 1999). In this way, different authors (e.g., Akter et al. 2013; Boyar et al. 2014) recognized that employing instruments at one point in time limits the ability to assess causal relationships. With the goal of remediating these issues and gaining a deeper understanding of the construct of interest, different studies (e.g., Morean et al. 2012; Schlosser and McNaughton 2009) suggest conducting a longitudinal study during the scale development. Using the longitudinal studies in this process may also allow the assessment of the scale’s predictive validity, since longitudinal designs evaluate whether the proposed interpretation of test scores can predict outcomes of interest over time. Therefore, future studies should consider the longitudinal approach in the scale development, both to facilitate greater understanding of the analyzed variables and to assess the predictive validity.

Self-reporting methodologies were also cited as limitations in some studies (e.g., Fisher et al. 2014; Pan et al. 2013). Mahudin et al. (2012) clarified that the self-reporting nature of quantitative studies raises the possibility of participant bias, social desirability, demand characteristics, and response sets. Such possibilities may, in turn, affect the validity of the findings. We agree with the authors’ suggestion that future research may also incorporate other objective or independent measures to supplement the subjective evaluation of the variables studied in the development of the new scale and to improve the interpretation of findings.

In addition, web-based surveys were another methodological limitation reported in some studies (e.g., Kim et al. 2011; Reed et al. 2011). Although this particular method has time- and cost-saving elements for data collection, its limitations are also highlighted. Researchers

have observed that important concerns include coverage bias (bias due to sampled individuals not having—or choosing not to access—the Internet) and nonresponse bias (bias due to participants of a survey differing from those who did not respond in terms of demographic or attitudinal variables) (Kim et al. 2011). Alternatives to minimize the problem in future research would be in-person surveys or survey interviews. Although more costly and more time consuming, these methods reduce problems related to concerns about confidentiality and the potential for coverage and nonresponse bias (Reed et al. 2011). Therefore, whenever possible, in-person surveys or survey interviews should be given priority in future research rather than web surveys.

Psychometric limitations Consistent with previous reports (MacKenzie et al. 2011; Prados 2007), this systematic review found distinct psychometric limitations reported in the scale development process. The lack of a more robust demonstration of construct validity and/or reliability was the most often mentioned limitation in the majority of the analyzed studies. For example, Alvarado-Herrera et al. (2015) reported the lack of a more robust demonstration of the predictive validity whereas Kim et al. (2011) of the nomological validity. Caro and Garcia (2007) noted that the relationships of the scale with other constructs were not analyzed. Saxena et al. (2015) and Pan et al. (2013) described the lack of demonstrable temporal stability (e.g., test-retest reliability). Imprecise or incomplete psychometric procedures that are employed during scale development are likely to obscure the outcome. Therefore, it is necessary for future research to consider adverse consequences for the reliability and validity of any construct, caused by poor test-theoretical practices. Only through detailed information and explanation of the rationale for statistical choices can the new measures be shown to have sufficient psychometric adjustments (Sveinbjornsdottir and Thorsteinsson 2008).

Additionally, the inadequate choice of the instruments or variables to be correlated with the variable of interest was another psychometric limitation cited in some studies (e.g., Bakar and Mustaffa 2013; Tanimura et al. 2011). This kind of limitation directly affects the convergent validity, which is a problem since, as has already been shown in this review, this type of validity has been one of the most recurrent practices in scale development. One hypothesis for this limitation may be the lack of gold standard measures to assess similar constructs as those of a new scale. In such cases, a relatively recent study by Morgado et al. (2014) offers a valid alternative. The authors used information collected on sociodemographic questionnaires (e.g., level of education and intensity of physical activity) to correlate with the

constructs of interest. Future researchers should seek support from the literature on the constructs that would be theoretically associated with the construct of interest, searching for alternatives in information collected on, for example, sociodemographic questionnaires, to assess the convergent validity of the new scale.

Another psychometric limitation reported in some studies was related to factor analysis. These limitations were identified in five main forms: (1) EFA and CFA were conducted using the data from the same sample (Zheng et al. 2010)—when this occurs, good model fit in the CFA is expected, as a consequence, the added strength of the CFA in testing a hypothesized structure for a new data set based on theory or previous findings is lost (Khine 2008); (2) lack of CFA (Bolton and Lane 2012)—if this happens, the researcher loses the possibility of assigning items to factors, testing the hypothesized structure of the data, and statistically comparing alternative models (Khine 2008); (3) a certain amount of subjectivity was necessary in identifying and labeling factors in EFA (Lombaerts et al. 2009)—since a factor is qualitative, it is common practice to label each factor based on an interpretation of the variables loading most heavily on it; the problem is that these labels are subjective in nature, represent the authors' interpretation, and can vary typically from 0.30 to 0.50 (Gottlieb et al. 2014; Khine 2008); (4) the initial unsatisfactory factor analysis output (Lombaerts et al. 2009); and (5) lack of a more robust CFA level (Jong et al. 2014) taken together—when the study result distances itself from statistical results expected for EFA (e.g., KMO, Bartlett test of sphericity) and/or CFA (e.g., CFI, GFI, RMSEA), it results in an important limitation, since the tested exploratory and theoretical models are not considered valid (Khine 2008). Taking these results, future studies should consider the use of separate samples for EFA and CFA, the combination of EFA and CFA, the definition of objective parameters to label factors, and about the consideration for unsatisfactory results of EFA and CFA, seeking alternatives to better fit the model.

Qualitative research limitations This review also found reported limitations on the qualitative approach of the analyzed studies. The first limitation was related to the exclusive use of the deductive method to generate items. It is noteworthy that, although most of the studies included in this review used exclusively deductive methods to generate items, only two studies recognized this as a limitation (Coleman et al. 2011; Song et al. 2011). Both studies used only the literature review to generate and operationalize the initial item pool. The authors recognized the importance of this deductive method to theoretically operationalize the target construct, but they noted that, “for further research, more diverse views

should be considered to reflect more comprehensive perspectives of human knowledge-creating behaviors to strengthen the validity of the developed scales” (Song et al. 2011, p. 256) and, “a qualitative stage could have been used to generate additional items [...]. This could also have reduced measurement error by using specific language the population used to communicate” (Coleman et al. 2011; p. 1069). Thus, the combination of deductive and inductive approaches (e.g., focus groups or interviews) in item generation is again suggested in future research.

In addition, it is also necessary that the researcher consider the quality of the reviewed literature. Napoli et al. (2014, p. 1096) reported limitations related to the loss of a more robust literature review, suggesting that the scale developed in the study may have been incorrectly operationalized: “Yet some question remains as to whether cultural symbolism should form part of this scale. Perhaps the way in which the construct was initially conceptualized and operationalized was incorrect.” The incorrect operation of the construct compromises the psychometric results of scale and its applicability in future studies.

Another limitation involves the subjective analysis of the qualitative research. Fisher et al. (2014, p. 488) pointed out that the qualitative methods (literature reviews and interviews) used to develop and conceptualize the construct were the main weaknesses of the study, “this research is limited by [...] the nature of qualitative research in which the interpretations of one researcher may not reflect those of another.” The authors explained that, due to the potential for researcher bias when interpreting data, it has been recognized that credible results are difficult to achieve. Nevertheless, subjective analysis is the essence and nature of qualitative studies. Some precautions in future studies can be taken to rule out potential researcher bias, such as attempts at neutrality. This is not always possible, however, and this limitation will remain a common problem in any qualitative study.

In turn, Sewitch et al. (2003, p. 260) reported that failure to formally assess content validity was a limitation. The reason given was budgetary constraints. It is worthwhile to remember that the content validity is an important step to ensure confidence in any inferences made using the final scale form. Therefore, it is necessarily required in any scale development process.

An additional limitation was reported by Lucas-Carrasco et al. (2011) in the recruitment of a larger number of interviewers, which may have affected the quality of the data collected. In order to minimize this limitation, the authors reported, “all interviewers had sufficient former education, received training on the study requirements, and were provided with a detailed guide” (p. 1223). Future studies

planning the use of multiple interviewers should consider potential resulting bias.

Missing data In connection, missing data was another issue reported by some studies included in this systematic review (e.g., Glynn et al. 2015; Ngorsuraches et al. 2007). Such limitations typically occur across different fields of scientific research. Missing data includes numbers that have been grouped, aggregated, rounded, censored, or truncated, resulting in partial loss of information (Schafer and Graham 2002). Collins et al. (2001) clarified that when researchers are confronted with missing data, they run an increased risk of reaching incorrect conclusions. This is because missing data may bias parameter estimates, inflate type I and type II error rates, and degrade the performance of confidence intervals. The authors also explained that, “because a loss of data is nearly always accompanied by a loss of information, missing values may dramatically reduce statistical power” (p. 330). Therefore, future researchers who wish to mitigate these risks during the scale development must pay close attention to the missing data aspect of the analysis and choose their strategy carefully.

Statistical methods to solve the problem of missing data have improved significantly, as demonstrated by Schafer and Graham (2002), although misconceptions still remain abundant. Several methods to deal with missing data were reviewed, issues raised, and advice offered for those that remain unresolved. Considering the fact that a more detailed discussion of the statistics dealing with missing data is beyond of the scope of this article, more details about missing data analysis can be found in Schafer and Graham (2002).

Social desirability bias Another limitation reported in some studies (Bova et al. 2006; Ngorsuraches et al. 2007) and identified in this systematic review is social desirability bias. This type of bias is considered to be a systematic error in self-reporting measures resulting from the desire of respondents to avoid embarrassment and project a favorable image to others (Fisher 1993). According to King and Bruner (2000), social desirability bias is an important threat to the validity of research employing multi-item scales. Provision of socially desirable responses in self-reported data may lead to spurious correlations between variables, as well as the suppression or moderation of relationships between the constructs of interest. Thus, one aspect of scale validity, which should be of particular concern to researchers, is the potential threat of contamination due to social-desirability response bias. To remedy this problem, we agree with the authors that it is incumbent upon researchers to identify situations in which data may be systematically biased toward the respondents’ perceptions of what is socially

acceptable, to determine the extent to which this represents contamination of the data, and to implement the most appropriate methods of control. Details on methods for identifying, testing for, and/or preventing social desirability bias are beyond the scope of this article, but can be found at King and Bruner (2000).

Item limitations In comparison with at least one previous study (Prados 2007), our findings reflect some potential item limitations. Firstly, items that were ambiguous or difficult to answer were the main weaknesses reported by Gottlieb et al. (2014). On this issue, the literature dealing with the necessary caution in wording the items is extensive. For example, items must clearly define the problem being addressed, must be as simple as possible, express a single idea, and use common words that reflect the vocabulary level of the target population. Items should not be inductors or have alternative or underlying assumptions. They must be free of generalizations and estimates, and be written to ensure the variability of responses. In writing the items, the researcher should avoid using fashionable expressions and colloquialisms or other words or phrases that impair understanding for groups of varying ages, ethnicities, religions, or genders. Furthermore, the items should be organized properly. For example, the opening questions should be simple and interesting to win the trust of the subjects. The most delicate, complex, or dull questions should be asked at the end of the sequence (Clark and Watson 1995; Malhotra 2004; Pasquali 2010).

Furthermore, Cicero et al. (2010) reported that the main limitation of their study was the fact that none of the items were reverse-scored. Although some methodologists claim that reverse scoring is necessary to avoid acquiescence among participants, this advice should be taken with caution. There are reports that the reverse-scored items may be confusing to participants, that the opposite of a construct reverse-scored may be fundamentally different than the construct, that reverse-scored items tend to be the worst fitting items in factor analyses, or that the factor structure of scales includes a factor with straightforward wording compared to a reverse-scored factor (Cicero et al. 2010). Awareness of these issues is necessary for future researchers to choose between avoiding acquiescence among participants or preventing a number of other problems related to the use of reverse scores.

Brevity of the scale Limitations on the scale size were also identified in this review. Studies by Negra and Mzoughi (2012) and Tombaugh et al. (2011) mentioned the short version of the scale as their main limitation. In both studies, the final version of the new scale included only five items. Generally, short scales are good, because

they require less time from respondents. However, very short scales can in fact seriously compromise the reliability of the instrument (Raykov 2008). To the extent that the researcher removes items of the scale, the Cronbach's alpha tends to decrease. It is valuable to remember that the minimum acceptable alpha should be at least 0.7, while an alpha value between 0.8 and 0.9 is considered ideal. Scales with many items tend to be more reliable, with higher alpha values (DeVellis 2003). In this context, future researchers should prioritize scales with enough items to keep the alpha within the acceptable range. Although many items may be lost during theoretical and psychometric analysis, an alternative already mentioned in this study would be to begin the initial item pool with at least twice the desired items of the final scale.

Difficulty controlling all variables In addition to all limitations reported, Gottlieb et al. (2014) mentioned a common limitation in different research fields—the difficulty of controlling all the variables that could influence the central construct of the study. The authors reported that “it may be that there are other variables that influence visitors’ perception of trade show effectiveness that were not uncovered in the research” and suggest “future research might yield insights that are not provided here” (p. 104). The reported limitation calls attention to the importance of the first step—item generation—in the scale development process. A possible remedy to this issue would be to know the target construct in detail during the item generation, allowing for all possible and important variables to be investigated and controlled. However, this is not always possible. Even using inductive and deductive approaches to generate items (literature review and interview), the authors still reported that limitation. In this light, future researchers must use care in hypothesizing and testing potential variables that could be controlled during construction of the scale development process.

Lack of manual instructions Finally, this review found a weakness reported on the loss of manualized instructions that regulate the data analysis. Saxena et al. (2015, p. 492) pointed out that the initial version of the new scale “did not contain manualized instructions for raters, so it lacked objective anchor points for choosing specific ratings on many of its questions”. Therefore, an important detail that should have the attention of future researchers are instructions that determine the application methods of the new scale. Pasquali (2010) suggests that when drafting the instructions, the researcher should define the development of operational strategies that will enable the application of the instrument and the format in which it will be presented and decide both how the

subject's response will be given for each item and the way that the respondent should answer each item. The researcher should also define how the scale scores would be analyzed. In addition, the instructions need to be as short as possible without confusion to the subjects of the target population, should contain one or more examples of how the items should be answered, and should ensure that the subject is free of any related tension or anxiety.

Study limitations and strengths

This review itself is subject to some limitations that should be taken into consideration. First, during the selection of the articles included in the analysis, we may have missed some studies that could have been identified by using other terms related to “scale development.” This may have impacted our findings. However, application of this term alone was recommended by its widespread use by researchers in the area (Clark and Watson 1995; DeVellis 2003; Hinkin 1995; Nunnally 1967) and by the large number of publications identified with this descriptor in the period evaluated, as compared with those screened with correlates (e.g., “development of questionnaire” and “development of measure”). In the same way, we may also have missed numerous studies that, despite recording their weaknesses, did not have the search term “limitations” indexed in the analyzed databases. We could have reduced this limitation by also using the search term ‘weakness’ or a similar word for selection and inclusion of several other articles. However, a larger number of included studies would hinder the operationalization of our findings.

Second, particularly regarding analysis of items and reliability, we lost information about the basic theories that support the scale development process: classical test theory (CTT)—known as classical psychometry—and item response theory (IRT)—known as modern psychometry (PASQUALI 2010). Although it was beyond the scope of this article to examine these theories, information on the employability of one or the other could contribute to a deeper understanding of their main limitations. Future studies could focus on CTT and IRT, compare the applicability of both, and identify their main limitations in the scale development process.

Still, our review is current with studies published until September 2015. As new evidence emerges on current practices and limitations reported in the scale development process, revisions to this systematic review and practice guideline would be required in future studies.

Despite its weaknesses, the strengths of this study should be highlighted. First, this study reviews the updated and consistent literature on scale development practices to be applied in, not only a specific field of knowledge as

carried out in most systematic review studies, but across various fields. With this variety of conceptions, we hope to assist future researchers in different areas of human and social sciences in making the most appropriate choice between strategies.

Second, this study differs from most studies of scale development revision, since it primarily considers the conceptions of the authors themselves about the main difficulties and mistakes made during the scale development process in their own studies. We hope to contribute to the efforts of future researchers, based on the knowledge of previous mistakes. While several weaknesses in scale development research were identified, specific recommendations for future research relevant to particular previously dimensions discussed were embedded within the appropriate sections throughout the article.

We observe that, although some weaknesses have been clearly identified in the scale development practices of many studies, only a few researchers recognized and recorded these limitations. This was evidenced in the large number of studies using exclusively deductive approaches to generate the initial item pool and the limited number of studies that recognized this as a limitation, or there were a large number of studies using smaller sample sizes than recommended in the literature for psychometric analysis and the limited number of studies that reported this issue as a limitation. Considering the observed distance between the limitation and its recognition, it is important that future researchers are comfortable with the detailed process of developing a new measure, especially as it pertains to avoiding theoretical and/or methodological mistakes, or at least, if they occur, to mention them as limitations.

Conclusions

In conclusion, the present research reviews numerous studies that both proposed current practices of the scale development process and also reported its main limitations. A variety of conceptions and methodological strategies and ten main limitations were identified and discussed along with suggestions for future research. In this way, we believe that this paper makes important contributions to the literature, especially because it provides a comprehensive set of recommendations to increase the quality of future practices in the scale development process.

Authors' contributions

FFRM is responsible for all parts of this manuscript, from its conception to the final writing. JFFM, CMN, ACSA and MECF participated in the data collection, analysis and interpretation of data and critical review of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interest.

Author details

¹Institute of Education, Universidade Federal Rural do Rio de Janeiro, BR-465, km 7, Seropédica, Rio de Janeiro 23890-000, Brazil. ²Faculty of Psychology, Universidade Federal de Juiz de Fora, Rua José Lourenço Kelmer, s/n—Campus Universitário Bairro São Pedro, Juiz de Fora, Minas Gerais 36036-900, Brazil. ³Faculty of Physical Education of the Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais, Av. Luz Interior, n 360, Estrela Sul, Juiz de Fora, Minas Gerais 36030-776, Brazil.

Received: 3 August 2016 Accepted: 22 December 2016

Published online: 25 January 2017

References

- Aagja, J. P., & Garg, R. (2010). Measuring perceived service quality for public hospitals (PubHosQual) in the Indian context. *International Journal of Pharmaceutical and Healthcare Marketing*, 4(10), 60–83. <http://dx.doi.org/10.1108/17506121011036033>.
- Ahmad, N., Awan, M. U., Raouf, A., & Sparks, L. (2009). Development of a service quality scale for pharmaceutical supply chains. *International Journal of Pharmaceutical and Healthcare Marketing*, 3(1), 26–45. <http://dx.doi.org/10.1108/17506120910948494>.
- Akter, S., D'Ambra, J., & Ray, P. (2013). Development and validation of an instrument to measure user perceived service quality of mHealth. *Information and Management*, 50, 181–195. <http://dx.doi.org/10.1016/j.im.2013.03.001>.
- Alvarado-Herrera, A., Bigne, E., Aldas-Manzano, J., & Curras-Perez, R. (2015). A scale for measuring consumer perceptions of corporate social responsibility following the sustainable development paradigm. *Journal of Business Ethics*, 1–20. doi: <http://dx.doi.org/10.1007/s10551-015-2654-9>
- Arias, M. R. M., Lloreda, M. J. H., & Lloreda, M. V. H. (2014). *Psicometría*. SA: Alianza Editorial
- Armfield, J. M. (2010). Development and psychometric evaluation of the Index of Dental Anxiety and Fear (IDAF-4C⁺). *Psychological Assessment*, 22(2), 279–287. <http://dx.doi.org/10.1037/a0018678>.
- Arrindell, W. A., & van der Ende, J. (1985). An empirical-test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement*, 9(2), 165–178. <http://dx.doi.org/10.1177/014662168500900205>.
- Atkins, K. G., & Kim, Y. (2012). Smart shopping: conceptualization and measurement. *International Journal of Retail and Distribution Management*, 40(5), 360–375. <http://dx.doi.org/10.1108/09590551211222349>.
- Bagdare, S., & Jain, R. (2013). Measuring retail customer experience. *International Journal of Retail and Distribution Management*, 41(10), 790–804. <http://dx.doi.org/10.1108/IJRD-08-2012-0084>.
- Bakar, H. A., & Mustafa, C. S. (2013). Organizational communication in Malaysia organizations. *Corporate Communications: An International Journal*, 18(1), 87–109. <http://dx.doi.org/10.1108/13563281311294146>.
- Barrett, P. T., & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality Study and Group Behavior*, 1, 23–33.
- Bastos, J. L., Celeste, R. K., Faerstein, E., & Barros, A. J. D. (2010). Racial discrimination and health: a systematic review of scales with a focus on their psychometric properties. *Social Science and Medicine*, 70, 1091–1099. <http://dx.doi.org/10.1016/j.socscimed.2009.12.20>.
- Beaudreuil, J., Allard, A., Zerkak, D., Gerber, RA, Cappelleri, JC, Quintero, N, Lasbleiz, S, ... Bardin, T. (2011). Unite' Rhumatologique des Affections de la Main (URAM) Scale: development and validation of a tool to assess Dupuytren's disease-specific disability. *Arthritis Care & Research*, 63(10), 1448–1455. doi: <http://dx.doi.org/10.1002/acr.20564>
- Bhattacharjee, A. (2002). Individual trust in online firms: scale development and initial test. *Journal of Management Information Systems*, 19(1), 211–241. <http://dx.doi.org/10.1080/07421222.2002.11045715>.
- Blankson, C., Cheng, J. M., & Spears, N. (2007). Determinants of banks selection in USA, Taiwan and Ghana. *International Journal of Bank Marketing*, 25(7), 469–489. <http://dx.doi.org/10.1108/02652320710832621>.
- Blankson, C., Paswan, A., & Boakye, K. G. (2012). College students' consumption of credit cards. *International Journal of Bank Marketing*, 30(7), 567–585. <http://dx.doi.org/10.1108/02652321211274327>.
- Bolton, D. L., & Lane, M. D. (2012). Individual entrepreneurial orientation: development of a measurement instrument. *Education + Training*, 54(2/3), 219–233. <http://dx.doi.org/10.1108/00400911211210314>.
- Bova, C., Fennie, K. P., Watrous, E., Dieckhaus, K., & Williams, A. B. (2006). The health care relationship (HCR) trust scale: development and psychometric evaluation. *Research in Nursing and Health*, 29, 477–488. <http://dx.doi.org/10.1002/nur.20158>.

- Bowen, H. P., & Wiersema, M. F. (1999). Matching method to paradigm in strategy research: limitations of cross-sectional analysis and some methodological alternatives. *Strategic Management Journal*, 20, 625–636.
- Boyar, S. L., Campbell, N. S., Mosley, D. C., Jr., & Carson, C. M. (2014). Development of a work/family social support measure. *Journal of Managerial Psychology*, 29(7), 901–920. <http://dx.doi.org/10.1108/JMP-06-2012-0189>.
- Brock, J. K., & Zhou, Y. (2005). Organizational use of the internet. *Internet Research*, 15(1), 67–87. <http://dx.doi.org/10.1108/10662240510577077>.
- Brun, I., Rajaobelina, L., & Ricard, L. (2014). Online relationship quality: scale development and initial testing. *International Journal of Bank Marketing*, 32(1), 5–27. <http://dx.doi.org/10.1108/IJBM-02-2013-0022>.
- Butt, M. M., & Run, E. C. (2010). Private healthcare quality: applying a SERVQUAL model. *International Journal of Health Care Quality Assurance*, 23(7), 658–673. <http://dx.doi.org/10.1108/09526861011071580>.
- Caro, L. M., & García, J. A. M. (2007). Measuring perceived service quality in urgent transport service. *Journal of Retailing and Consumer Services*, 14, 60–72. <http://dx.doi.org/10.1016/j.jretconser.2006.04.001>.
- Chahal, H., & Kumari, N. (2012). Consumer perceived value. *International Journal of Pharmaceutical and Healthcare Marketing*, 6(2), 167–190. <http://dx.doi.org/10.1108/17506121211243086>.
- Chen, H., Tian, Y., & Daugherty, P. J. (2009). Measuring process orientation. *The International Journal of Logistics Management*, 20(2), 213–227. <http://dx.doi.org/10.1108/09574090910981305>.
- Choi, S. W., Victorson, D. E., Yount, S., Anton, S., & Cella, D. (2011). Development of a conceptual framework and calibrated item banks to measure patient-reported dyspnea severity and related functional limitations. *Value in Health*, 14, 291–306. <http://dx.doi.org/10.1016/j.jval.2010.06.001>.
- Christophersen, T., & Konradt, U. (2012). Development and validation of a formative and a reflective measure for the assessment of online store usability. *Behaviour and Information Technology*, 31(9), 839–857. <http://dx.doi.org/10.1080/0144929X.2010.529165>.
- Churchill, G. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16(1), 64–73. <http://dx.doi.org/10.2307/3150876>.
- Cicero, D. C., Kerns, J. G., & McCarthy, D. M. (2010). The Aberrant Salience Inventory: a new measure of psychosis proneness. *Psychological Assessment*, 22(3), 688–701. <http://dx.doi.org/10.1037/a0019913>.
- Clark, L. A., & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, 7(3), 309–319. <http://dx.doi.org/10.1037/1040-3590.7.3.309>.
- Coker, B. L. S., Ashill, N. J., & Hope, B. (2011). Measuring internet product purchase risk. *European Journal of Marketing*, 45(7/8), 1130–1151. <http://dx.doi.org/10.1108/03090561111137642>.
- Coleman, D., Chernatony, L., & Christodoulides, G. (2011). B2B service brand identity: scale development and validation. *Industrial Marketing Management*, 40, 1063–1071. <http://dx.doi.org/10.1016/j.indmarman.2011.09.010>.
- Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330–351. <http://dx.doi.org/10.1037/1082-989X.6.4.330>.
- Colwell, S. R., Aung, M., Kanetkar, V., & Holden, A. L. (2008). Toward a measure of service convenience: multiple-item scale development and empirical test. *Journal of Services Marketing*, 22(2), 160–169. <http://dx.doi.org/10.1108/08876040810862895>.
- Cossette, S., Cara, C., Ricard, N., & Pepin, J. (2005). Assessing nurse–patient interactions from a caring perspective: report of the development and preliminary psychometric testing of the Caring Nurse–Patient Interactions Scale. *International Journal of Nursing Studies*, 42, 673–686. <http://dx.doi.org/10.1016/j.ijnurstu.2004.10.004>.
- Dennis, R. S., & Bocarnea, M. (2005). Development of the servant leadership assessment instrument. *Leadership and Organization Development Journal*, 26(8), 600–615. <http://dx.doi.org/10.1108/01437730510633692>.
- DeVellis, R. F. (2003). *Scale development: theory and applications* (2nd ed.). Newbury Park: Sage Publications.
- Devlin, J. F., Roy, S. K., & Sekhon, H. (2014). Perceptions of fair treatment in financial services. *European Journal of Marketing*, 48(7/8), 1315–1332. <http://dx.doi.org/10.1108/EJM-08-2012-0469>.
- Dunham, A., & Burt, C. (2014). Understanding employee knowledge: the development of an organizational memory scale. *The Learning Organization*, 21(2), 126–145. <http://dx.doi.org/10.1108/TLO-04-2011-0026>.
- Edwards, J. R., Knight, D. K., Broome, K. M., & Flynn, P. M. (2010). The development and validation of a transformational leadership survey for substance use treatment programs. *Substance Use and Misuse*, 45, 1279–1302. <http://dx.doi.org/10.3109/10826081003682834>.
- Feuerstein, M., Nicholas, R. A., Huang, G. D., Haufler, A. J., Pransky, G., & Robertson, M. (2005). Workstyle: development of a measure of response to work in those with upper extremity pain. *Journal of Occupational Rehabilitation*, 15(2), 87–104. <http://dx.doi.org/10.1007/s10926-005-3420-0>.
- Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, 20(2), 303–315. <http://dx.doi.org/10.1086/209351>.
- Fisher, R., Maritz, A., & Lobo, A. (2014). Evaluating entrepreneurs' perception of success. *International Journal of Entrepreneurial Behavior and Research*, 20(5), 478–492. <http://dx.doi.org/10.1108/IJEBR-10-2013-0157>.
- Flight, R. L., D'Souza, G., & Allaway, A. W. (2011). Characteristics-based innovation adoption: scale and model validation. *Journal of Product and Brand Management*, 20(5), 343–355. <http://dx.doi.org/10.1108/10610421111157874>.
- Forbush, K. T., Wildes, J. E., Pollack, L. O., Dunbar, D., Luo, J., Patterson, P., Petruzzelli, L., ... Watson, D. (2013). Development and validation of the Eating Pathology Symptoms Inventory (EPSI). *Psychological Assessment*, 25(3), 859–878. doi: <http://dx.doi.org/10.1037/a0032639>.
- Foster, J. D., McCain, J. L., Hibberts, M. F., Brunell, A. B., & Johnson, B. (2015). The grandiose narcissism scale: a global and facet-level measure of grandiose narcissism. *Personality and Individual Differences*, 73, 12–16. <http://dx.doi.org/10.1016/j.paid.2014.08.042>.
- Franché, R., Corbière, M., Lee, H., Breslin, F. C., & Hepburn, G. (2007). The readiness for return-to-work (RRTW) scale: development and validation of a self-report staging scale in lost-time claimants with musculoskeletal disorders. *Journal of Occupational Rehabilitation*, 17, 450–472. <http://dx.doi.org/10.1007/s10926-007-9097-9>.
- Gesten, E. L. (1976). A health resources inventory: the development of a measure of the personal and social competence of primary-grade children. *Journal of Consulting and Clinical Psychology*, 44(5), 775–786. <http://dx.doi.org/10.1037/0022-006X.44.5.775>.
- Gibbons, C. J., Kenning, C., Coventry, P. A., Bee, P., Bundy, C., Fisher, L., & Bower, P. (2013). Development of a Multimorbidity Illness Perceptions Scale (MULTIPLEs). *PLoS One*, 8(12), e81852. <http://dx.doi.org/10.1371/journal.pone.0081852>.
- Gligor, D. M., & Holcomb, M. (2014). The road to supply chain agility: an RBV perspective on the role of logistics capabilities. *The International Journal of Logistics Management*, 25(1), 160–179. <http://dx.doi.org/10.1108/IJLM-07-2012-0062>.
- Glynn, N. W., Santanasto, A. J., Simonsick, E. M., Boudreau, R. M., Beach, S. R., Schulz, R., & Newman, A. B. (2015). The Pittsburgh fatigability scale for older adults: development and validation. *Journal of American Geriatrics Society*, 63, 130–135. <http://dx.doi.org/10.1111/jgs.13191>.
- Gottlieb, U., Brown, M., & Ferrier, L. (2014). Consumer perceptions of trade show effectiveness. *European Journal of Marketing*, 48(1/2), 89–107. <http://dx.doi.org/10.1108/EJM-06-2011-0310>.
- Hair Junior, J. F., Black, W. C., Babin, N. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados* (6th ed.). São Paulo: Bookman.
- Hall, M. A., Camacho, F., Dugan, E., & Balkrishnan, R. (2002). Trust in the medical profession: conceptual and measurement issues. *Health Services Research*, 37(5), 1419–1439. <http://dx.doi.org/10.1111/1475-6773.01070>.
- Han, H., Back, K., & Kim, Y. (2011). A multidimensional scale of switching barriers in the full-service restaurant industry. *Cornell Hospitality Quarterly*, 52(1), 54–63. <http://dx.doi.org/10.1177/1938965510389261>.
- Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research*, 57, 98–107. [http://dx.doi.org/10.1016/S0148-2963\(01\)00295-8](http://dx.doi.org/10.1016/S0148-2963(01)00295-8).
- Henderson-King, D., & Henderson-King, E. (2005). Acceptance of cosmetic surgery: scale development and validation. *Body Image*, 2, 137–149. <http://dx.doi.org/10.1016/j.bodyim.2005.03.003>.
- Hernandez, J. M. C., & Santos, C. C. (2010). Development-based trust: proposing and validating a new trust measurement model for buyer-seller relationships. *Brazilian Administration Review*, 7(2), 172–197. <http://dx.doi.org/10.1590/S1807-76922010000200005>.
- Hildebrandt, T., Langenbacher, J., & Schlundt, D. G. (2004). Muscularity concerns among men: development of attitudinal and perceptual measures. *Body Image*, 1, 169–181. <http://dx.doi.org/10.1016/j.bodyim.2004.01.001>.
- Hilsenroth, M. J., Blagys, M. D., Ackerman, S. J., Bonge, D. R., & Blais, M. A. (2005). Measuring psychodynamic-interpersonal and cognitive-behavioral techniques: development of the comparative psychotherapy process scale. *Psychotherapy: Theory, Research, Practice, Training*, 42(3), 340–356. <http://dx.doi.org/10.1037/0033-3204.42.3.340>.

- Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, 21(5), 967–988. <http://dx.doi.org/10.1177/014920639502100509>.
- Ho, C. B., & Lin, W. (2010). Measuring the service quality of internet banking: scale development and validation. *European Business Review*, 22(1), 5–24. <http://dx.doi.org/10.1108/09555341011008981>.
- Hutz, C.S., Bandeira, D.R., & Trentini, C.M. (Org.). (2015). *Psicometria*. Porto Alegre, Artmed.
- Jong, N., Van Leeuwen, R. G. J., Hoekstra, H. A., & van der Zee, K. I. (2014). CRIQ: an innovative measure using comparison awareness to avoid self-presentation tactics. *Journal of Vocational Behavior*, 84, 199–214. <http://dx.doi.org/10.1016/j.jvb.2014.01.003>.
- Kapuscinski, A. N., & Masters, K. S. (2010). The current status of measures of spirituality: a critical review of scale development. *Psychology of Religion and Spirituality*, 2(4), 191–205. <http://dx.doi.org/10.1037/a0020498>.
- Khine, M. S. (2008). *Knowing, knowledge and beliefs: epistemological studies across diverse cultures*. New York: Springer.
- Khorsan, R., & Crawford, C. (2014). External validity and model validity: a conceptual approach for systematic review methodology. *Evidence-Based Complementary and Alternative Medicine*, 2014, Article ID 694804, 12 pages. doi: <http://dx.doi.org/10.1155/2014/694804>
- Kim, S., Cha, J., Knutson, B. J., & Beck, J. A. (2011). Development and testing of the Consumer Experience Index (CEI). *Managing Service Quality: An International Journal*, 21(2), 112–132. <http://dx.doi.org/10.1108/0960452111113429>.
- Kim, D., Lee, Y., Lee, J., Nam, J. K., & Chung, Y. (2014). Development of Korean smartphone addiction proneness scale for youth. *PLoS One*, 9(5), e97920. <http://dx.doi.org/10.1371/journal.pone.0097920>.
- King, M. F., & Bruner, G. C. (2000). Social desirability bias: a neglected aspect of validity testing. *Psychology and Marketing*, 17(2), 79–103. [http://dx.doi.org/10.1002/\(SICI\)1520-6793\(200002\)17:2<79::AID-MAR2>3.0.CO;2-0](http://dx.doi.org/10.1002/(SICI)1520-6793(200002)17:2<79::AID-MAR2>3.0.CO;2-0).
- Kwon, W., & Lennon, S. J. (2011). Assessing college women's associations of American specialty apparel brands. *Journal of Fashion Marketing and Management: An International Journal*, 15(2), 242–256. <http://dx.doi.org/10.1108/13612021111132663>.
- Ladhari, R. (2010). Developing e-service quality scales: a literature review. *Journal of Retailing and Consumer Services*, 17, 464–477. <http://dx.doi.org/10.1016/j.jretconser.2010.06.003>.
- Lin, J. C., & Hsieh, P. (2011). Assessing the self-service technology encounters: development and validation of SSTQUAL scale. *Journal of Retailing*, 87(2), 194–206. <http://dx.doi.org/10.1016/j.jretail.2011.02.006>.
- Lombaerts, K., Backer, F., Engels, N., Van Braak, J., & Athanasou, J. (2009). Development of the self-regulated learning teacher belief scale. *European Journal of Psychology of Education*, 24(1), 79–96. <http://dx.doi.org/10.1007/BF03173476>.
- Lucas-Carrasco, R., Eser, E., Hao, Y., McPherson, K. M., Green, A., & Kullmann, L. (2011). The quality of care and support (QOCS) for people with disability scale: development and psychometric properties. *Research in Developmental Disabilities*, 32, 1212–1225. <http://dx.doi.org/10.1016/j.ridd.2010.12.030>.
- MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: integrating new and existing techniques. *MIS Quarterly*, 35(2), 293–334.
- Mahudin, N. D. M., Cox, T., & Griffiths, A. (2012). Measuring rail passenger crowding: scale development and psychometric properties. *Transportation Research Part, F*, 15, 38–51. <http://dx.doi.org/10.1016/j.trf.2011.11.006>.
- Malhotra, N. K. (2004). *Pesquisa de marketing: Uma orientação aplicada* (4th ed.). Porto Alegre: Bookman.
- Medina-Pradas, C., Navarro, J. B., López, S. R., Grau, A., & Obiols, J. E. (2011). Further development of a scale of perceived expressed emotion and its evaluation in a sample of patients with eating disorders. *Psychiatry Research*, 190, 291–296. <http://dx.doi.org/10.1016/j.psychres.2011.06.011>.
- Meneses, J., Barrios, M., Bonillo, A., Cosculluela, A., Lozano, L. M., Turbany, J., & Valero, S. (2014). *Psicometria*. Barcelona: Editorial UOC.
- Morean, M. E., Corbin, W. R., & Treat, T. A. (2012). The anticipated effects of alcohol scale: development and psychometric evaluation of a novel assessment tool for measuring alcohol expectancies. *Psychological Assessment*, 24(4), 1008–1023. <http://dx.doi.org/10.1037/a0028982>.
- Morgado, F. F. R., Campana, A. N. N. B., & Tavares, M. C. G. C. F. (2014). Development and validation of the self-acceptance scale for persons with early blindness: the SAS-EB. *PLoS One*, 9(9), e106848. <http://dx.doi.org/10.1371/journal.pone.0106848>.
- Nagy, B. G., Blair, E. S., & Lohrke, F. T. (2014). Developing a scale to measure liabilities and assets of newness after start-up. *International Entrepreneurship and Management Journal*, 10, 277–295. <http://dx.doi.org/10.1007/s11365-012-0219-2>.
- Napoli, J., Dickinson, S. J., Beverland, M. B., & Farrelly, F. (2014). Measuring consumer-based brand authenticity. *Journal of Business Research*, 67, 1090–1098. <http://dx.doi.org/10.1016/j.jbusres.2013.06.001>.
- Negra, A., & Mzoughi, M. N. (2012). How wise are online procrastinators? A scale development. *Internet Research*, 22(4), 426–442. <http://dx.doi.org/10.1108/10662241211250971>.
- Ngorsuraches, S., Lerkiatbundit, S., Li, S. C., Treesak, C., Sirithorn, R., & Korwiwattanakarn, M. (2007). Development and validation of the patient trust in community pharmacists (TRUST-Ph) scale: results from a study conducted in Thailand. *Research in Social and Administrative Pharmacy*, 4, 272–283. <http://dx.doi.org/10.1016/j.sapharm.2007.10.002>.
- Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw Hill.
- Oh, H. (2005). Measuring affective reactions to print apparel advertisements: a scale development. *Journal of Fashion Marketing and Management: An International Journal*, 9(3), 283–305. <http://dx.doi.org/10.1108/13612020510610426>.
- Olaya, B., Marsà, F., Ochoa, S., Balanzá-Martínez, V., Barbeito, S., González-Pinto, A., ... Haro, J.M. (2012). Development of the insight scale for affective disorders (ISAD): modification from the scale to assess unawareness of mental disorder. *Journal of Affective Disorders*, 142, 65–71. doi: <http://dx.doi.org/10.1016/j.jad.2012.03.041>.
- Omar, N. A., & Musa, R. (2011). Measuring service quality in retail loyalty programmes (LPSQual). *International Journal of Retail and Distribution Management*, 39(10), 759–784. <http://dx.doi.org/10.1108/09590551111162257>.
- Pan, J., Wong, D. F. K., & Ye, S. (2013). Post-migration growth scale for Chinese international students: development and validation. *Journal of Happiness Studies*, 14, 1639–1655. <http://dx.doi.org/10.1007/s10902-012-9401-z>.
- Pasquali, L. (2010). *Instrumentação psicológica: fundamentos e práticas*. Porto Alegre: Artmed.
- Patwardhan, H., & Balasubramanian, S. K. (2011). Brand romance: a complementary approach to explain emotional attachment toward brands. *Journal of Product and Brand Management*, 20(4), 297–308. <http://dx.doi.org/10.1108/10610421111148315>.
- Pimentel, C. E., Gouveia, V. V., & Pessoa, V. S. (2007). Escala de Preferência Musical: construção e comprovação da sua estrutura fatorial. *Psico-USF*, 12(2), 145–155.
- Podsakoff, N. P., Podsakoff, P. M., MacKenzie, S. B., & Klinger, R. L. (2013). Are we really measuring what we say we're measuring? Using video techniques to supplement traditional construct validation procedures. *Journal of Applied Psychology*, 98(1), 99–113. <http://dx.doi.org/10.1037/a0029570>.
- Pommer, A.M., Prins, L., van Ranst, D., Meijer, J., Hul, A.V., Janssen, J., ... Pop, V.J.M. (2013). Development and validity of the Patient-Centred COPD Questionnaire (PCQ). *Journal of Psychosomatic Research*, 75, 563–571. doi: <http://dx.doi.org/10.1016/j.jpsychores.2013.10.001>
- Prados, J. M. (2007). Development of a new scale of beliefs about the worry consequences. *Annals of Psychology*, 23(2), 226–230.
- Raykov, T. (2008). Alpha if item deleted: a note on loss of criterion validity in scale development if maximizing coefficient alpha. *British Journal of Mathematical and Statistical Psychology*, 61, 275–285. <http://dx.doi.org/10.1348/000711007X188520>.
- Reed, L. L., Vidaver-Cohen, D., & Colwell, S. R. (2011). A new scale to measure executive servant leadership: development, analysis, and implications for research. *Journal of Business Ethics*, 101, 415–434. <http://dx.doi.org/10.1007/s10551-010-0729-1>.
- Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, 12(3), 287–297. <http://dx.doi.org/10.1037/1040-3590.12.3.287>.
- Rice, S. M., Fallon, B. J., Aucote, H. M., & Möller-Leimkühler, A. M. (2013). Development and preliminary validation of the male depression risk scale: Furthering the assessment of depression in men. *Journal of Affective Disorders*, 151, 950–958. <http://dx.doi.org/10.1016/j.jad.2013.08.013>.
- Riedel, M., Spellmann, I., Schennach-Wolff, R., Obermeier, M., & Musil, R. (2011). The RSM-scale: a pilot study on a new specific scale for self- and observer-rated quality of life in patients with schizophrenia. *Quality of Life Research*, 20, 263–272. <http://dx.doi.org/10.1007/s11136-010-9744-z>.
- Roberson, R. B., III, Elliott, T. R., Chang, J. E., & Hill, J. N. (2014). Exploratory factor analysis in rehabilitation psychology: a content analysis. *Rehabilitation Psychology*, 59(4), 429–438. <http://dx.doi.org/10.1037/a0037899>.
- Rodrigues, A. C. A., & Bastos, A. V. B. (2012). Organizational entrenchment: scale development and validation. *Psicologia: Reflexão e Crítica*, 25(4), 688–700. <http://dx.doi.org/10.1590/S0102-79722012000400008>.
- Rodríguez, I., Kozusznik, M. W., & Peiró, J. M. (2013). Development and validation of the Valencia Eustress-Distress Appraisal Scale. *International Journal of Stress Management*, 20(4), 279–308. <http://dx.doi.org/10.1037/a0034330>.

- Rosenthal, S. (2011). Measuring knowledge of indoor environmental hazards. *Journal of Environmental Psychology*, 31, 137–146. <http://dx.doi.org/10.1016/j.jenvp.2010.08.003>.
- Saxena, S., Ayers, C. R., Dozier, M. E., & Maidment, K. M. (2015). The UCLA Hoarding Severity Scale: development and validation. *Journal of Affective Disorders*, 175, 488–493. <http://dx.doi.org/10.1016/j.jad.2015.01.030>.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the Art. *Psychological Methods*, 7(2), 147–177. <http://dx.doi.org/10.1037//1082-989X.7.2.147>.
- Schlosser, F. K., & McNaughton, R. B. (2009). Using the I-MARKOR scale to identify market-oriented individuals in the financial services sector. *Journal of Services Marketing*, 23(4), 236–248. <http://dx.doi.org/10.1108/08876040910965575>.
- Sewitch, M. J., Abrahamowicz, M., Dobkin, P. L., & Tamblyn, R. (2003). Measuring differences between patients' and physicians' health perceptions: the patient–physician discordance scale. *Journal of Behavioral Medicine*, 26(3), 245–263. <http://dx.doi.org/10.1023/A:1023412604715>.
- Sharma, P. (2010). Measuring personal cultural orientations: scale development and validation. *Journal of the Academy of Marketing Science*, 38, 787–806. <http://dx.doi.org/10.1007/s11747-009-0184-7>.
- Sharma, D., & Gassenheimer, J. B. (2009). Internet channel and perceived cannibalization. *European Journal of Marketing*, 43(7/8), 1076–1091. <http://dx.doi.org/10.1108/03090560910961524>.
- Shawyer, F., Ratcliff, K., Mackinnon, A., Farhall, J., Hayes, S. C., & Copolov, D. (2007). The Voices Acceptance and Action Scale (VAAS): pilot data. *Journal of Clinical Psychology*, 63(6), 593–606. <http://dx.doi.org/10.1002/jclp.20366>.
- Sin, L. Y. M., Tse, A. C. B., & Yim, F. H. K. (2005). CRM: conceptualization and scale development. *European Journal of Marketing*, 39(11/12), 1264–1290. <http://dx.doi.org/10.1108/03090560510623253>.
- Sohn, D., & Choi, S. M. (2014). Measuring expected interactivity: scale development and validation. *New Media and Society*, 16(5), 856–870. <http://dx.doi.org/10.1177/1461444813495808>.
- Song, J. H., Uhm, D., & Yoon, S. W. (2011). Organizational knowledge creation practice. *Leadership and Organization Development Journal*, 32(3), 243–259. <http://dx.doi.org/10.1108/01437731111123906>.
- Staines, Z. (2013). Managing tacit investigative knowledge: measuring "investigative thinking styles". *Policing: An International Journal of Police Strategies and Management*, 36(3), 604–619. <http://dx.doi.org/10.1108/PIJPSM-07-2012-0072>.
- Sultan, P., & Wong, H. (2010). Performance-based service quality model: an empirical study on Japanese universities. *Quality Assurance in Education*, 18(2), 126–143. <http://dx.doi.org/10.1108/09684881011035349>.
- Sveinbjornsdottir, S., & Thorsteinsson, E. B. (2008). Adolescent coping scales: a critical psychometric review. *Scandinavian Journal of Psychology*, 49(6), 533–548. <http://dx.doi.org/10.1111/j.1467-9450.2008.00669.x>.
- Swaid, S. I., & Wigand, R. T. (2009). Measuring the quality of E-Service: scale development and initial validation. *Journal of Electronic Commerce Research*, 10(1), 13–28.
- Tanimura, C., Morimoto, M., Hiramatsu, K., & Hagino, H. (2011). Difficulties in the daily life of patients with osteoarthritis of the knee: scale development and descriptive study. *Journal of Clinical Nursing*, 20, 743–753. <http://dx.doi.org/10.1111/j.1365-2702.2010.03536.x>.
- Taute, H. A., & Sierra, J. (2014). Brand tribalism: an anthropological perspective. *Journal of Product and Brand Management*, 23(1), 2–15. <http://dx.doi.org/10.1108/JPB-06-2013-0340>.
- Tombaugh, J. R., Mayfield, C., & Durand, R. (2011). Spiritual expression at work: exploring the active voice of workplace spirituality. *International Journal of Organizational Analysis*, 19(2), 146–170. <http://dx.doi.org/10.1108/19348831111135083>.
- Turker, D. (2009). Measuring corporate social responsibility: a scale development study. *Journal of Business Ethics*, 85, 411–427. <http://dx.doi.org/10.1007/s10551-008-9780-6>.
- Uzunboylu, H., & Ozdamli, F. (2011). Teacher perception for m-learning: scale development and teachers' perceptions. *Journal of Computer Assisted Learning*, 27, 544–556. <http://dx.doi.org/10.1111/j.1365-2729.2011.00415.x>.
- Van der Gaag, M., Schütz, C., ten Napel, A., Landa, Y., Delespaul, P., Bak, M., ... Hert, M. (2013). Development of the Davos Assessment of Cognitive Biases Scale (DACOBS). *Schizophrenia Research*, 144, 63–71. doi: <http://dx.doi.org/10.1016/j.schres.2012.12.010>
- Von Steinbüchel, N., Wilson, L., Gibbons, H., Hawthorne, G., Höfer, S., Schmidt, S., ... Truelle, J. (2010). *Journal of Neurotrauma*, 27, 1167–1185. doi: <http://dx.doi.org/10.1089/neu.2009.1076>
- Voon, B. H., Abdullah, F., Lee, N., & Kueh, K. (2014). Developing a HospiSE scale for hospital service excellence. *International Journal of Quality and Reliability Management*, 31(3), 261–280. <http://dx.doi.org/10.1108/IJQRM-10-2012-0143>.
- Walshe, M., Peach, R. K., & Miller, N. (2009). Dysarthria Impact Profile: development of a scale to measure psychosocial effects. *International Journal of Language and Communication Disorders*, 44(5), 693–715. <http://dx.doi.org/10.1080/13682820802317536>.
- Wang, C. L., & Mowen, J. C. (1997). The separateness-connectedness self-schema: scale development and application to message construction. *Psychology and Marketing*, 14(2), 185–207. [http://dx.doi.org/10.1002/\(SICI\)1520-6793\(199703\)14:2<185::AID-MAR5>3.0.CO;2-9](http://dx.doi.org/10.1002/(SICI)1520-6793(199703)14:2<185::AID-MAR5>3.0.CO;2-9).
- Wepener, M., & Boshoff, C. (2015). An instrument to measure the customer-based corporate reputation of large service organizations. *Journal of Services Marketing*, 29(3), 163–172. <http://dx.doi.org/10.1108/JSM-01-2014-0026>.
- Williams, Z., Ponder, N., & Autry, C. W. (2009). Supply chain security culture: measure development and validation. *The International Journal of Logistics Management*, 20(2), 243–260. <http://dx.doi.org/10.1108/09574090910981323>.
- Wilson, N. L., & Holmvall, C. M. (2013). The development and validation of the incivility from customers scale. *Journal of Occupational Health Psychology*, 18(3), 310–326. <http://dx.doi.org/10.1037/a0032753>.
- Yang, M., Weng, S., & Hsiao, P. (2014). Measuring blog service innovation in social media services. *Internet Research*, 24(1), 110–128. <http://dx.doi.org/10.1108/IntR-12-2012-0253>.
- Zhang, X., & Hu, D. (2011). Farmer-buyer relationships in China: the effects of contracts, trust and market environment. *China Agricultural Economic Review*, 3(1), 42–53. <http://dx.doi.org/10.1108/17561371111103534>.
- Zheng, J., You, L., Lou, T., Chen, N., Lai, D., Liang, Y., ... Zhai, C. (2010). Development and psychometric evaluation of the dialysis patient-perceived exercise benefits and barriers scale. *International Journal of Nursing Studies*, 47, 166–180. doi: <http://dx.doi.org/10.1016/j.ijnurstu.2009.05.023>

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com

Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine

ABSTRACT

The task of developing a new questionnaire or translating an existing questionnaire into a different language might be overwhelming. The greatest challenge perhaps is to come up with a questionnaire that is psychometrically sound, and is efficient and effective for use in research and clinical settings. This article provides guidelines for the development and translation of questionnaires for application in medical fields, with a special emphasis on perioperative and pain medicine. We provide a framework to guide researchers through the various stages of questionnaire development and translation. To ensure that the questionnaires are psychometrically sound, we present a number of statistical methods to assess the reliability and validity of the questionnaires.

Key words: Anesthesia; development; questionnaires; translation; validation

Introduction

Questionnaires or surveys are widely used in perioperative and pain medicine research to collect quantitative information from both patients and health-care professionals. Data of interest could range from observable information (e.g., presence of lesion, mobility) to patients' subjective feelings of their current status (e.g., the amount of pain they feel, psychological status). Although using an existing questionnaire will save time and resources,^[1] a questionnaire that measures the construct of interest may not be readily available, or the published questionnaire is not available in the language required for the targeted respondents. As a result, investigators may need to develop a new questionnaire or translate an existing one

into the language of the intended respondents. Prior work has highlighted the wealth of literature available on psychometric principles, methodological concepts, and techniques regarding questionnaire development/translation and validation. To that end, this article is not meant to provide an exhaustive review of all the related statistical concepts and methods. Rather, this article aims to provide straightforward guidelines for the development or translation of questionnaires (or scales) for use in perioperative and pain medicine research for readers who may be unfamiliar with the process of questionnaire development and/or translation. Readers are recommended to consult the cited references to further examine these techniques for application.

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as the author is credited and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

How to cite this article: Tsang S, Royse CF, Terkawi AS. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. Saudi J Anaesth 2017;11:S80-9.

Access this article online

Website:
www.saudija.org

DOI:
10.4103/sja.SJA_203_17

Quick Response Code



SINY TSANG, COLIN F. ROYSE^{1,2}, ABDULLAH SULIEMAN TERKAWI^{3,4,5}

Department of Epidemiology, Columbia University, New York, NY, ³Department of Anesthesiology, University of Virginia, Charlottesville, VA, ⁵Outcomes Research Consortium, Cleveland, OH, USA, ¹Department of Surgery, University of Melbourne, Melbourne, ²Department of Anesthesia and Pain Management, The Royal Melbourne Hospital, Parkville, Victoria, Australia, ⁴Department of Anesthesiology, King Fahad Medical City, Riyadh, Saudi Arabia

Address for correspondence: Dr. Siny Tsang, Department of Epidemiology, Columbia University, New York, NY, USA.
E-mail: st2989@cumc.columbia.edu

This article is divided into two main sections. The first discusses issues that investigators should be aware of in developing or translating a questionnaire. The second section of this paper illustrates procedures to validate the questionnaire after the questionnaire is developed or translated. A model for the questionnaire development and translation process is presented in Figure 1. In this special issue of the Saudi Journal of Anesthesia we presented multiple studies of development and validation of questionnaires in perioperative and pain medicine, we encourage readers to refer to them for practical experience.

Preliminary Considerations

It is crucial to identify the construct that is to be assessed with the questionnaire, as the domain of interest will determine what the questionnaire will measure. The next question is: How will the construct be operationalized? In other words, what types

of behavior will be indicative of the domain of interest? Several approaches have been suggested to help with this process,^[2] such as content analysis, review of research, critical incidents, direct observations, expert judgment, and instruction.

Once the construct of interest has been determined, it is important to conduct a literature review to identify if a previously validated questionnaire exists. A validated questionnaire refers to a questionnaire/scale that has been developed to be administered among the intended respondents. The validation processes should have been completed using a representative sample, demonstrating adequate reliability and validity. Examples of necessary validation processes can be found in the validation section of this paper. If no existing questionnaires are available, or none that are determined to be appropriate, it is appropriate to construct a new questionnaire. If a questionnaire exists, but only in a different language, the task is to translate and validate the questionnaire in the new language.

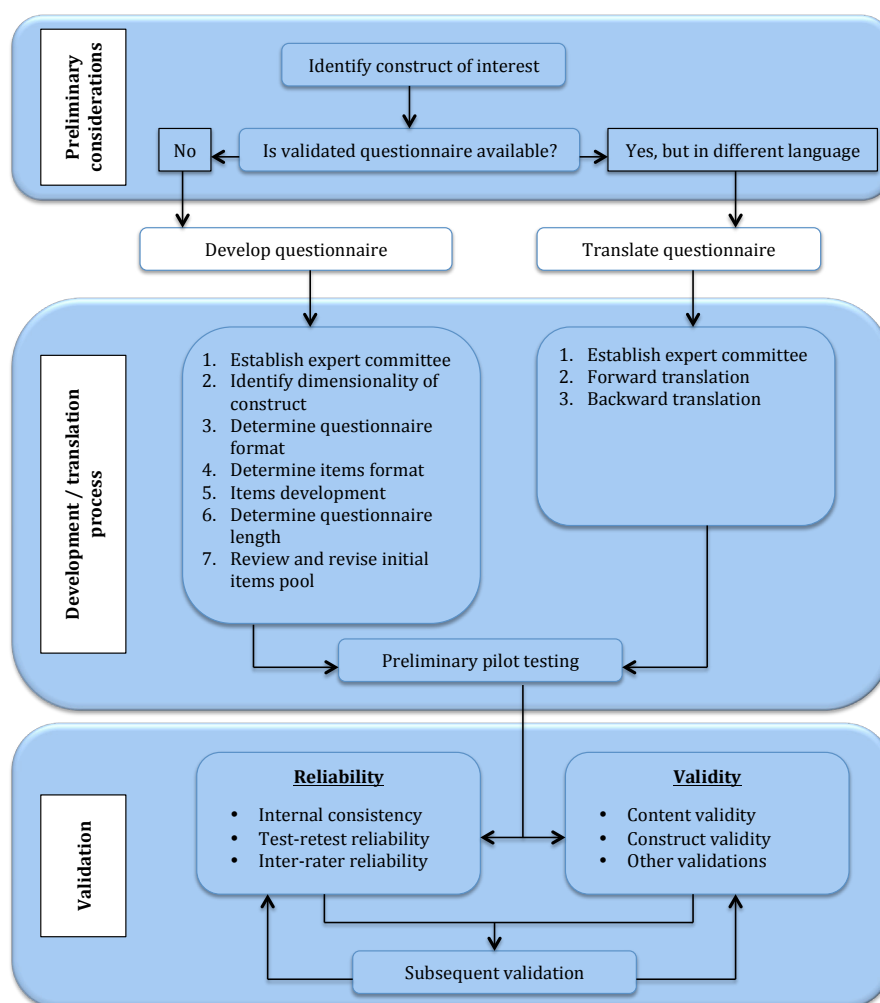


Figure 1: Questionnaire development and translation processes

Developing a Questionnaire

To construct a new questionnaire, a number of issues should be considered even before writing the questionnaire items.

Identify the dimensionality of the construct

Many constructs are multidimensional, meaning that they are composed of several related components. To fully assess the construct, one may consider developing subscales to assess the different components of the construct. Next, are all the dimensions equally important? or are some more important than others? If the dimensions are equally important, one can assign the same weight to the questions (e.g., by summing or taking the average of all the items). If some dimensions are more important than others, it may not be reasonable to assign the same weight to the questions. Rather, one may consider examining the results from each dimension separately.

Determine the format in which the questionnaire will be administered

Will the questionnaire be self-administered or administered by a research/clinical staff? This decision depends, in part, on what the questionnaire intends to measure. If the questionnaire is designed to measure catastrophic thinking related to pain, respondents may be less likely to respond truthfully if a research/clinical staff asked the questions, whereas they may be more likely to respond truthfully if they are allowed to complete the questionnaire on their own. If the questionnaire is designed to measure patients' mobility after surgery, respondents may be more likely to overreport the amount of mobility in an effort to demonstrate recovery. To obtain a more accurate measure of mobility after surgery, it may be preferable to obtain objective ratings by clinical staff.

If respondents are to complete the questionnaire by themselves, the items need to be written in a way that can be easily understood by the majority of the respondents, generally about Grade 6 reading level.^[3] If the questionnaire is to be administered to young respondents or respondents with cognitive impairment, the readability level of the items should be lowered. Questionnaires intended for children should take into consideration the cognitive stages of young people^[4] (e.g., pictorial response choices may be more appropriate, such as pain faces to assess pain^[5]).

Determine the item format

Will the items be open ended or close ended? Questions that are open ended allow respondents to elaborate upon their responses. As more detailed information may be obtained using open-ended questions, these items are best suited for situations in which investigators wish to gather more information about a specific domain. However, these

responses are often more difficult to code and score, which increases the difficulty of summarizing individuals' responses. If multiple coders are included, researchers have to address the additional issue of inter-rater reliability.

Questions that are close ended provide respondents a limited number of response options. Compared to open-ended questions, these items are easier to administer and analyze. On the other hand, respondents may not be able to clarify their responses, and their responses may be influenced by the response options provided.

If close-ended items are to be used, should multiple-choice, Likert-type scales, true/false, or other close-ended formats be used? How many response options should be available? If a Likert-type scale is to be adopted, what scale anchors are to be used to indicate the degree of agreement (e.g., strongly agree, agree, neither, disagree, strongly disagree), frequency of an event (e.g., almost never, once in a while, sometimes, often, almost always), or other varying options? To make use of participants' responses for subsequent statistical analyses, researchers should keep in mind that items should be scaled to generate sufficient variance among the intended respondents.^[6,7]

Item development

A number of guidelines have been suggested for writing items.^[7] Items should be simple, short, and written in language familiar to the target respondents. The perspective should be consistent across items; items that assess affective responses (e.g., anxiety, depression) should not be mixed with those that assess behavior (e.g., mobility, cognitive functioning).^[8] Items should assess only a single issue. Items that address more than one issue, or "double-barreled" items (e.g., "My daily activities and mood are affected by my pain."), should not be used. Avoid leading questions as they may result in biased responses. Items that all participants would respond similarly (e.g., "I would like to reduce my pain.") should not be used, as the small variance generated will provide limited information about the construct being assessed. Table 1 summarizes important tips on writing questions.

The issue of whether reverse-scored items should be used remains debatable. Since reverse-scored items are negatively worded, it has been argued that the inclusion of these items may reduce response set bias.^[9] On the other hand, others have found a negative impact on the psychometric properties of scales that included negatively worded items.^[10] In recent years, an increasing amount of literature reports problems with reverse-scored items.^[11-14] Researchers who decide to include negatively worded items should take extra steps

Table 1: Tips on writing questions^[15,16]

Use short and simple sentences
Ask for only one piece of information at a time. Example: Have you had nausea and vomiting in the last 24 h? Someone may have nausea, but may not have vomiting, thus this question should be divided into two questions
Avoid negatives if possible. Example: In the last 24 h, how many times did you not have pruritus? The better format would be; in the last 24 h, how many times did you have pruritus?
Ask precise questions. Example: Have you had pain before? Better question would be; what was your worst pain in the last 24 h?
Ensure that those you ask have the necessary knowledge. Example: Have you had neuropathic pain before? Many patients may not know what "neuropathic" means, a better question(s) would be to ask about the symptoms of neuropathic pain, for example, "have you had episodes of piercing pain like hot needles into your skin, before"
Avoid unnecessary details, as people are usually less inclined to complete long questionnaires, however make sure to ask for all the essential details
Avoid asking direct questions on sensitive issues. Example: "Are you obese?" can be better written as "do you think you have a weight issue"
Minimize bias. Example: "I was satisfied with the pain management that I had (yes or no)?" Better question is to ask about the level of satisfaction in a scale from 0 to 10. As many patients may choose yes to please you
Avoid weasel words such as commonly, usually, some, and hardly ever. Example: "Do you commonly have pain?" is better written as "How often do you have pain?"
Avoid using statements instead of questions
Avoid using agreement response anchors. Example: Your postoperative pain was the main concern to you before surgery (with Likert scale options). A better question would be what was your main concern before surgery? (with listing some options)
Avoid using too few or too many response anchors. Use five or more response anchors to achieve stable participant responses
Verbally label each response option, use only verbal labels, maintain equal spacing between response options, and use additional space to visually separate nonsubstantive response options from the substantive options
Arrange the questions. Always go from general to particular, easy to difficult, and factual to abstract
Consider adding some contradictory questions, to detect the responders' consistency, as some tend to tick whether "agree" or "disagree"

to ensure that the items are interpreted as intended by the respondents, and that the reverse-coded items have similar psychometric properties as the other regularly coded items.^[7]

Determine the intended length of questionnaire

There is no rule of thumb for the number of items that make up a questionnaire. The questionnaire should contain sufficient items to measure the construct of interest, but not be so long that respondents experience fatigue or loss of motivation in completing the questionnaire.^[17,18] Not only should a questionnaire possess the most parsimonious (i.e., simplest) structure,^[19] but it also should consist of items that adequately represent the construct of interest to minimize measurement error.^[20] Although a simple structure of questionnaire is recommended, a large pool of items is needed in the early stages of the questionnaire's development as many of these items might be discarded throughout the development process.^[7]

Review and revise initial pool of items

After the initial pool of questionnaire items are written, qualified experts should review the items. Specifically, the items should be reviewed to make sure they are accurate, free of item construction problems, and grammatically correct. The reviewers should, to the best of their ability, ensure that the items do not contain content that may be perceived as offensive or biased by a particular subgroup of respondents.

Preliminary pilot testing

Before conducting a pilot test of the questionnaire on the intended respondents, it is advisable to test the questionnaire items on a small sample (about 30–50)^[21] of

respondents.^[17] This is an opportunity for the questionnaire developer to know if there is confusion about any items, and whether respondents have suggestions for possible improvements of the items. One can also get a rough idea of the response distribution to each item, which can be informative in determining whether there is enough variation in the response to justify moving forward with a large-scale pilot test. Feasibility and the presence of floor (almost all respondents scored near the bottom) or ceiling effects (almost all respondents scored near the top) are important determinants of items that are included or rejected at this stage. Although it is possible that participants' responses to questionnaires may be affected by question order,^[22-24] this issue should be addressed only after the initial questionnaire has been validated. The questionnaire items should be revised upon reviewing the results of the preliminary pilot testing. This process may be repeated a few times before finalizing the final draft of the questionnaire.

Summary

So far, we highlighted the major steps that need to be undertaken when constructing a new questionnaire. Researchers should be able to clearly link the questionnaire items to the theoretical construct they intend to assess. Although such associations may be obvious to researchers who are familiar with the specific topic, they may not be apparent to other readers and reviewers. To develop a questionnaire with good psychometric properties that can subsequently be applied in research or clinical practice, it is crucial to invest the time and effort to ensure that the items adequately assess the construct of interest.

Translating a Questionnaire

The following section summarizes the guidelines for translating a questionnaire into a different language.

Forward translation

The initial translation from the original language to the target language should be made by at least two independent translators.^[25,26] Preferably, the bilingual translators should be translating the questionnaire into their mother tongue, to better reflect the nuances of the target language.^[27] It is recommended that one translator be aware of the concepts the questionnaire intend to measure, to provide a translation that more closely resembles the original instrument. It is suggested that a naïve translator, who is unaware of the objective of the questionnaire, produce the second translation so that subtle differences in the original questionnaire may be detected.^[25,26] Discrepancies between the two (or more) translators can be discussed and resolved between the original translators, or with the addition of an unbiased, bilingual translator who was not involved in the previous translations.

Backward translation

The initial translation should be independently back-translated (i.e., translate back from the target language into the original language) to ensure the accuracy of the translation. Misunderstandings or unclear wordings in the initial translations may be revealed in the back-translation.^[25] As with the forward translation, the backward translation should be performed by at least two independent translators, preferably translating into their mother language (the original language).^[26] To avoid bias, back-translators should preferably not be aware of the intended concepts the questionnaire measures.^[25]

Expert committee

Constituting an expert committee is suggested to produce the prefinal version of the translation.^[25] Members of the committee should include experts who are familiar with the construct of interest, a methodologist, both the forward and backward translators, and if possible, developers of the original questionnaires. The expert committee will need to review all versions of the translations and determine whether the translated and original versions achieve semantic, idiomatic, experiential, and conceptual equivalence.^[25,28] Any discrepancies will need to be resolved, and members of the expert committee will need to reach a consensus on all items to produce a prefinal version of the translated questionnaire. If necessary, the process of translation and back-translation can be repeated.

Preliminary pilot testing

As with developing a new questionnaire, the prefinal version of the translated questionnaire should be pilot tested on a small sample (about 30–50)^[21] of the intended respondents.^[25,26] After completing the translated questionnaire, the respondent is asked (verbally by an interviewer or via an open-ended question) to elaborate what they thought each questionnaire item and their corresponding response meant. This approach allows the investigator to make sure that the translated items retained the same meaning as the original items, and to ensure there is no confusion regarding the translated questionnaire. This process may be repeated a few times to finalize the final translated version of the questionnaire.

Summary

In this section, we provided a template for translating an existing questionnaire into a different language. Considering that most questionnaires were initially developed in one language (e.g., English when developed in English-speaking countries^[25]), translated versions of the questionnaires are needed for researchers who intend to collect data among respondents who speak other languages. To compare responses across populations of different language and/or culture, researchers need to make sure that the questionnaires in different languages are assessing the equivalent construct with an equivalent metric. Although the translation process is time consuming and costly, it is the best method to ensure that a translated measure is equivalent to the original questionnaire.^[28]

Validating a Questionnaire

Initial validation

After the new or translated questionnaire items pass through preliminary pilot testing and subsequent revisions, it is time to conduct a pilot test among the intended respondents for initial validation. In this pilot test, the final version of the questionnaire is administered to a large representative sample of respondents for whom the questionnaire is intended. If the pilot test is conducted for small samples, the relatively large sampling errors may reduce the statistical power needed to validate the questionnaire.^[2]

Reliability

The reliability of a questionnaire can be considered as the consistency of the survey results. As measurement error is present in content sampling, changes in respondents, and differences across raters, the consistency of a questionnaire can be evaluated using its internal consistency, test-retest reliability, and inter-rater reliability, respectively.

Internal consistency

Internal consistency reflects the extent to which the questionnaire items are inter-correlated, or whether they are consistent in measurement of the same construct. Internal consistency is commonly estimated using the coefficient alpha,^[29] also known as Cronbach's alpha. Given a questionnaire x , with k number of items, alpha (α) can be computed as:

$$\alpha = \frac{\kappa}{\kappa - 1} \left(1 - \frac{\sum \sigma_i^2}{\sigma_x^2} \right)$$

Where, σ_i^2 is the variance of item i , and σ_x^2 is the total variance of the questionnaire.

Cronbach's alpha ranges from 0 to 1 (when some items are negatively correlated with other items in the questionnaire, it is possible to have negative values of Cronbach's alpha). When reverse-scored items are [incorrectly] not reverse scored, it can be easily remedied by correctly scoring the items. However, if a negative Cronbach's alpha is still obtained when all items are correctly scored, there are serious problems in the original design of the questionnaire), with higher values indicating that items are more strongly interrelated with one another. Cronbach's $\alpha = 0$ indicates no internal consistency (i.e., none of the items are correlated with one another), whereas $\alpha = 1$ reflects perfect internal consistency (i.e., all the items are perfectly correlated with one another). In practice, Cronbach's alpha of at least 0.70 has been suggested to indicate adequate internal consistency.^[30] A low Cronbach's alpha value may be due to poor inter-relatedness between items; as such, items with low correlations with the questionnaire total score should be discarded or revised. As alpha is a function of the length of the questionnaire, alpha will increase with the number of items. In addition, alpha will increase if the variability of each item is increased. It is, therefore, possible to increase alpha by including more related items, or adding items that have more variability to the questionnaire. On the other hand, an alpha value that is too high ($\alpha \geq 0.90$) suggests that some questionnaire items may be redundant;^[31] investigators may consider removing items that are essentially asking the same thing in multiple ways.

It is important to note that Cronbach's alpha is a property of the responses from a specific sample of respondents.^[31] Investigators need to keep in mind that Cronbach's alpha is not "the" estimate of reliability for a questionnaire under all circumstances. Rather, the alpha value only indicates the extent to which the questionnaire is reliable for "a particular population of examinees."^[32] A questionnaire with excellent reliability with one sample may not necessarily

have the same reliability in another. Therefore, the reliability of a questionnaire should be estimated each time the questionnaire is administered, including pilot testing and subsequent validation stages.

Test-retest reliability

Test-retest reliability refers to the extent to which individuals' responses to the questionnaire items remain relatively consistent across repeated administration of the same questionnaire or alternate questionnaire forms.^[2] Provided the same individuals were administered the same questionnaires twice (or more), test-retest reliability can be evaluated using Pearson's product moment correlation coefficient (Pearson's r) or the intraclass correlation coefficient.

Pearson's r between the two questionnaires' responses can be referred to as the coefficient of stability. A larger stability coefficient indicates stronger test-retest reliability, reflecting that measurement error of the questionnaire is less likely to be attributable to changes in the individuals' responses over time.

Test-retest reliability can be considered the stability of respondents' attributes; it is applicable to questionnaires that are designed to measure personality traits, interest, or attitudes that are relatively stable across time, such as anxiety and pain catastrophizing. If the questionnaires are constructed to measure transitory attributes, such as pain intensity and quality of recovery, test-retest reliability is not applicable as the changes in respondents' responses between assessments are reflected in the instability of their responses. Although test-retest reliability is sometimes reported for scales that are intended to assess constructs that change between administrations, researchers should be aware that test-retest reliability is not applicable and does not provide useful information about the questionnaires of interest. Researchers should also be critical when evaluating the reliability estimates reported in such studies.

An important question to consider in estimating test-retest reliability is how much time should lapse between questionnaire administrations? If the duration between time 1 and time 2 is too short, individuals may remember their responses in time 1, which may overestimate the test-retest reliability. Respondents, especially those recovering from major surgery, may experience fatigue if the retest is administered shortly after the first administration, which may underestimate the test-retest reliability. On the other hand, if there is a long period of time between questionnaire administrations, individuals' responses may change due to other factors (e.g., a respondent may be taking pain

management medications to treat chronic pain condition). Unfortunately, there is no single answer. The duration should be long enough to allow the effects of memory to fade and to prevent fatigue, but not so long as to allow changes to take place that may affect the test-retest reliability estimate.^[17]

Inter-rater reliability

For questionnaires in which multiple raters complete the same instrument for each examinee (e.g., a checklist of behavior/symptoms), the extent to which raters are consistent in their observations across the same group of examinees can be evaluated. This consistency is referred to as the inter-rater reliability, or inter-rater agreement, and can be estimated using the kappa statistic.^[33] Suppose two clinicians independently rated the same group of patients on their mobility after surgery (e.g., 0 = needs help of 2+ people; 1 = needs help of 1 person; 2 = independent), kappa (κ) can be computed as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Where, P_o is the observed proportion of observations in which the two raters agree, and P_e is the expected proportion of observations in which the two raters agree by chance. Accordingly, κ is the proportion of agreement between the two raters, after factoring out the proportion of agreement by chance. κ ranges from 0 to 1, where $\kappa = 0$ indicates all chance agreements and $\kappa = 1$ represents perfect agreement between the two raters. Others have suggested $\kappa = 0$ as no agreement, $\kappa = 0.01 - 0.20$ as poor agreement, $\kappa = 0.21 - 0.40$ as slight agreement, $\kappa = 0.41 - 0.60$ as fair agreement, $\kappa = 0.61 - 0.80$ as good agreement, $\kappa = 0.81 - 0.92$ as very good agreement, and $\kappa = 0.93 - 1$ as excellent agreement.^[34,35] If more than two raters are used, an extension of Cohen's κ statistic is available to compute the inter-rater reliability across multiple raters.^[36]

Validity

The validity of a questionnaire is determined by analyzing whether the questionnaire measures what it is intended to measure. In other words, are the inferences and conclusions made based on the results of the questionnaire (i.e., test scores) valid?^[37] Two major types of validity should be considered when validating a questionnaire: content validity and construct validity.

Content validity

Content validity refers to the extent to which the items in a questionnaire are representative of the entire theoretical construct the questionnaire is designed to assess.^[17] Although the construct of interest determines which items are written

and/or selected in the questionnaire development/translation phase, content validity of the questionnaire should be evaluated after the initial form of the questionnaire is available.^[2] The process of content validation is particularly crucial in the development of a new questionnaire.

A panel of experts who are familiar with the construct that the questionnaire is designed to measure should be tasked with evaluating the content validity of the questionnaire. The experts judge, as a panel, whether the questionnaire items are adequately measuring the construct intended to assess, and whether the items are sufficient to measure the domain of interest. Several approaches to quantify the judgment of content validity across experts are also available, such as the content validity ratio^[38] and content validation form.^[39,40] Nonetheless, as the process of content validation depends heavily on how well the panel of experts can assess the extent to which the construct of interest is operationalized, the selection of appropriate experts is crucial to ensure that content validity is evaluated adequately. Example items to assess content validity include:^[41]

- The questions were clear and easy
- The questions covered all the problem areas with your pain
- You would like the use of this questionnaire for future assessments
- The questionnaire lacks important questions regarding your pain
- Some of the questions violate your privacy.

A concept that is related to content validity is face validity. Face validity refers to the degree to which the respondents or laypersons judge the questionnaire items to be valid. Such judgment is based less on the technical components of the questionnaire items, but rather on whether the items appear to be measuring a construct that is meaningful to the respondents. Although this is the weakest way to establish the validity of a questionnaire, face validity may motivate respondents to answer more truthfully. For example, if patients perceive a quality of recovery questionnaire to be evaluating how well they are recovering from surgery, they may be more likely to respond in ways that reflect their recovery status.

Construct validity

Construct validity is the most important concept in evaluating a questionnaire that is designed to measure a construct that is not directly observable (e.g., pain, quality of recovery). If a questionnaire lacks construct validity, it will be difficult to interpret results from the questionnaire, and inferences cannot be drawn from questionnaire responses to a

behavior domain. The construct validity of a questionnaire can be evaluated by estimating its association with other variables (or measures of a construct) with which it should be correlated positively, negatively, or not at all.^[42] In practice, the questionnaire of interest, as well as the preexisting instruments that measure similar and dissimilar constructs, is administered to the same groups of individuals. Correlation matrices are then used to examine the expected patterns of associations between different measures of the same construct, and those between a questionnaire of a construct and other constructs. It has been suggested that correlation coefficients of 0.1 should be considered as small, 0.3 as moderate, and 0.5 as large.^[43]

For instance, suppose a new scale is developed to assess pain among hospitalized patients. To provide evidence of construct validity for this new pain scale, we can examine how well patients' responses on the new scale correlate with the preexisting instruments that also measure pain. This is referred to as convergent validity. One would expect strong correlations between the new questionnaire and the existing measures of the same construct, since they are measuring the same theoretical construct.

Alternatively, the extent to which patients' responses on the new pain scale correlate with instruments that measure unrelated constructs, such as mobility or cognitive function, can be assessed. This is referred to as divergent validity. As pain is theoretically dissimilar to the constructs of mobility or cognitive function, we would expect zero, or very weak, correlation between the new pain questionnaire and

instruments that assess mobility or cognitive function. Table 2 describes different validation types and important definitions.

Subsequent validation

The process described so far defines the steps for initial validation. However, the usefulness of the scale is the ability to discriminate between different cohorts in the domain of interest. It is advised that several studies investigating different cohorts or interventions should be conducted to identify whether the scale can discriminate between groups. Ideally, these studies should have clearly defined outcomes where the changes in the domain of interest are well known. For example, in subsequent validation of the Postoperative Quality of Recovery Scale, four studies were constructed to show the ability to discriminate recovery and cognition in different cohorts of participants (mixed cohort, orthopedics, and otolaryngology), as well as a human volunteer study to calibrate the cognitive domain.^[46-49]

Sample size

Guidelines for the respondent-to-item ratio ranged from 5:1^[50] (i.e., fifty respondents for a 10-item questionnaire), 10:1,^[30] to 15:1 or 30:1.^[51] Others suggested that sample sizes of 50 should be considered as very poor, 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1000 or more as excellent.^[52] Given the variation in the types of questionnaire being used, there are no absolute rules for the sample size needed to validate a questionnaire.^[53] As larger samples are always better than smaller samples, it is recommended that investigators utilize as large a sample size as possible. The respondent-to-item ratios can be utilized

Table 2: Questionnaire-related terminology^[16,44,45]

Terminology	Definitions
Construct	A model, idea, or theory that the researcher is attempting to assess (e.g., quality of postoperative recovery)
Validity	The ability of a questionnaire to truly measure what it purports to measure
Reliability	Reliability or reproducibility is the ability of a questionnaire to produce the same results when administered at two different points of time
Content validity	The extent to which a questionnaire measure includes the most relevant and important aspects of a concept in the context of a given measurement application
Face validity	The ability of an instrument to be understandable and relevant to the targeted population
Construct validity	The degree to which scores on the questionnaire measure relate to other measures (e.g., patient reported or clinical indicators) in a manner that is consistent with theoretically derived <i>a priori</i> hypotheses concerning the concepts that are being measured
Diagnostic validity	The accuracy of a questionnaire in diagnosing certain conditions (e.g., neuropathic pain)
Known-group validity	The ability of a questionnaire to be sensitive to differences between groups of patients that may be anticipated to score differently in the predicted direction
Criterion validity	The ability of a questionnaire to measure how well one measure predicts an outcome for another measure
Concurrent validity	The association of an instrument with accepted standards
Predictive validity	The ability of a questionnaire to predict future health status or test results. Future health status is considered a better indicator than the true value or a standard
Internal consistency	The degree of the inter-relatedness among the items in a multi-item questionnaire measure. It is usually measured by Cronbach's alpha
Repeatability (test-retest reliability)	The ability of the scores of an instrument to be reproducible if it is used on the same patient while the patient's condition has not changed (measurements repeated over time)
Responsiveness	The extent to which a questionnaire measure can detect changes in the construct being measured over time. It is applicable only for questionnaires that are designed to assess changes in the construct within a short period of time

to further strengthen the rationale for the large sample size when necessary.

Other considerations

Even though data collection using questionnaires is relatively easy, researchers should be cognizant about the necessary approvals that should be obtained prior to beginning the research project. Considering the differences in regulations and requirements in different countries, agencies, and institutions, researchers are advised to consult the research ethics committee at their agencies and/or institutions regarding the necessary approval needed and additional considerations that should be addressed.

Conclusion

In this review, we provided guidelines on how to develop, validate, and translate a questionnaire for use in perioperative and pain medicine. The development and translation of a questionnaire requires investigators' thorough consideration of issues relating to the format of the questionnaire and the meaning and appropriateness of the items. Once the development or translation stage is completed, it is important to conduct a pilot test to ensure that the items can be understood and correctly interpreted by the intended respondents. The validation stage is crucial to ensure that the questionnaire is psychometrically sound. Although developing and translating a questionnaire is no easy task, the processes outlined in this article should enable researchers to end up with questionnaires that are efficient and effective in the target populations.

Financial support and sponsorship

Siny Tsang, PhD, was supported by the research training grant 5-T32-MH 13043 from the National Institute of Mental Health.

Conflicts of interest

There are no conflicts of interest.

References

- Boynton PM, Greenhalgh T. Selecting, designing, and developing your questionnaire. *BMJ* 2004;328:1312-5.
- Crocker L, Algina J. Introduction to Classical and Modern Test Theory. Mason, Ohio: Cengage Learning; 2008.
- Davis TC, Mayeaux EJ, Fredrickson D, Bocchini JA Jr., Jackson RH, Murphy PW. Reading ability of parents compared with reading level of pediatric patient education materials. *Pediatrics* 1994;93:460-8.
- Bell A. Designing and testing questionnaires for children. *J Res Nurs* 2007;12:461-9.
- Wong DL, Baker CM. Pain in children: Comparison of assessment scales. *Okla Nurse* 1988;33:8.
- Stone E. Research Methods in Organizational Behavior. Glenview, IL: Scott Foresman; 1978.
- Hinkin TR. A brief tutorial on the development of measures for use in survey questionnaires. *Organ Res Methods* 1998;2:104-21.
- Harrison DA, McLaughlin ME. Cognitive processes in self-report responses: Tests of item context effects in work attitude measures. *J Appl Psychol* 1993;78:129-40.
- Price JL, Mueller CW. Handbook of Organizational Measurement. Marshfield, MA: Pitman; 1986.
- Harrison DA, McLaughlin ME. Exploring the Cognitive Processes Underlying Responses to Self-Report Instruments: Effects of Item Content on Work Attitude Measures. *Academy of Management Annual Meetings*; 1991. p. 310-4.
- Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers; 2000.
- Lindwall M, Barkoukis V, Grano C, Lucidi F, Raudsepp L, Liukkonen J, *et al.* Method effects: The problem with negatively versus positively keyed items. *J Pers Assess* 2012;94:196-204.
- Stansbury JP, Ried LD, Velozo CA. Unidimensionality and bandwidth in the Center for Epidemiologic Studies Depression (CES-D) Scale. *J Pers Assess* 2006;86:10-22.
- Tsang S, Salekin RT, Coffey CA, Cox J. A comparison of self-report measures of psychopathy among non-forensic samples using item response theory analyses. *Psychol Assess*. [In press].
- Leung WC. How to design a questionnaire. *Stud BMJ* 2001;9.
- Artino AR Jr., La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No 87. *Med Teach* 2014;36:463-74.
- Schultz KS, Whitney DJ. Measurement Theory in Action: Case Studies and Exercises. Thousand Oaks, CA: Sage; 2005.
- Schmitt NW, Stults DM. Factors defined by negatively keyed items: The results of careless respondents? *Appl Psychol Meas* 1985;9:367-73.
- Thurstone LL. Multiple-Factor Analysis. Chicago, IL: University of Chicago Press; 1947.
- Churchill GA. A paradigm for developing better measures of marketing constructs. *J Mark Res* 1979;16:64-73.
- Perneger TV, Courvoisier DS, Hudelson PM, Gayet-Ageron A. Sample size for pre-tests of questionnaires. *Qual Life Res* 2015;24:147-51.
- Bowling A, Windsor J. The effects of question order and response-choice on self-rated health status in the English Longitudinal Study of Ageing (ELSA). *J Epidemiol Community Health* 2008;62:81-5.
- Lee S, Schwarz N. Question context and priming meaning of health: Effect on differences in self-rated health between Hispanics and non-Hispanic Whites. *Am J Public Health* 2014;104:179-85.
- Schwarz N. Self-reports: How the questions shape the answers. *Am Psychol* 1999;54:93-105.
- Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *J Clin Epidemiol* 1993;46:1417-32.
- Beaton D, Bombardier C, Guillemin F, Ferraz M. Recommendations for the Cross-Cultural Adaptation of the DASH and Quick DASH Outcome Measures. Toronto: Institute for Work and Health; 2007.
- Hendricson WD, Russell IJ, Prihoda TJ, Jacobson JM, Rogan A, Bishop GD, *et al.* Development and initial validation of a dual-language English-Spanish format for the Arthritis Impact Measurement Scales. *Arthritis Rheum* 1989;32:1153-9.
- Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila Pa 1976)* 2000;25:3186-91.
- Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297-334.
- Nunnally J. Psychometric Theory. New York: McGraw-Hill; 1978.
- Streiner DL. Starting at the beginning: An introduction to coefficient alpha and internal consistency. *J Pers Assess* 2003;80:99-103.
- Wilkinson L, the Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. *Am Psychol*

- 1999;54:594-604.
33. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37-46.
 34. Dawson B, Trapp RG. *Basic and Clinical Biostatistics*. 3rd ed. Norwalk, Conn.: Lange Medical Books; 2001.
 35. Grootsholten C, Bajema IM, Florquin S, Steenbergen EJ, Peutz-Kootstra CJ, Goldschmeding R, *et al.* Inter-observer agreement of scoring of histopathological characteristics and classification of lupus nephritis. *Nephrol Dial Transplant* 2008;23:223-30.
 36. Berry KJ, Mielke PW. A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educ Psychol Meas* 1988;48:921-33.
 37. Murphy KR, Davidshofer CO. *Psychological Testing: Principles and Applications*. Upper Saddle River, NJ: Prentice Hall; 2001.
 38. Lawshe CH. A quantitative approach to content validity. *Pers Psychol* 1975;28:563-75.
 39. Barrett RS. Content validation form. *Public Pers Manage* 1992;21:41-52.
 40. Barrett RS, editor. Content validation form. In: *Fair Employment Strategies in Human Resource Management*. Westport, CT: Quorum Books/Greenwood; 1996. p. 47-56.
 41. Alnahhal A, May S. Validation of the arabic version of the quebec back pain disability Scale. *Spine (Phila Pa 1976)* 2012;37:E1645-50.
 42. Cronbach L, Meehl P. Construct validity in psychological tests. *Psychol Bull* 1955;52:281-302.
 43. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Erlbaum; 1988.
 44. Anthoine E, Moret L, Regnault A, Sébille V, Hardouin JB. Sample size used to validate a scale: A review of publications on newly-developed patient reported outcomes measures. *Health Qual Life Outcomes* 2014;12:176.
 45. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, *et al.* ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res* 2013;22:1889-905.
 46. Newman S, Wilkinson DJ, Royse CF. Assessment of early cognitive recovery after surgery using the Post-operative Quality of Recovery Scale. *Acta Anaesthesiol Scand* 2014;58:185-91.
 47. Royse CF, Newman S, Williams Z, Wilkinson DJ. A human volunteer study to identify variability in performance in the cognitive domain of the postoperative quality of recovery scale. *Anesthesiology* 2013;119:576-81.
 48. Royse CF, Williams Z, Purser S, Newman S. Recovery after nasal surgery vs. tonsillectomy: Discriminant validation of the Postoperative Quality of Recovery Scale. *Acta Anaesthesiol Scand* 2014;58:345-51.
 49. Royse CF, Williams Z, Ye G, Wilkinson D, De Steiger R, Richardson M, *et al.* Knee surgery recovery: Post-operative Quality of Recovery Scale comparison of age and complexity of surgery. *Acta Anaesthesiol Scand* 2014;58:660-7.
 50. Gorsuch RL. *Factor Analysis*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1983.
 51. Pedhazur RJ. *Multiple Regression in Behavioral Research: Explanation and Prediction*. Fort Worth, TX: Harcourt Brace College Publishers; 1997.
 52. Comfrey AL, Lee HB. *A First Course in Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1992.
 53. Osborne JW, Costello AB. Sample size and subject to item ratio in principal components analysis. *Pract Assess Res Eval* 2004;9:8.



Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer

Godfred O. Boateng^{1*}, Torsten B. Neilands², Edward A. Frongillo³,
Hugo R. Melgar-Quinonez⁴ and Sera L. Young^{1,5}

¹ Department of Anthropology and Global Health, Northwestern University, Evanston, IL, United States, ² Division of Prevention Science, Department of Medicine, University of California, San Francisco, San Francisco, CA, United States, ³ Department of Health Promotion, Education and Behavior, Arnold School of Public Health, University of South Carolina, Columbia, SC, United States, ⁴ Institute for Global Food Security, School of Human Nutrition, McGill University, Montreal, QC, Canada, ⁵ Institute for Policy Research, Northwestern University, Evanston, IL, United States

OPEN ACCESS

Edited by:

Jimmy Thomas Efrid,
University of Newcastle, Australia

Reviewed by:

Aida Turrini,
Consiglio per la Ricerca in Agricoltura
e L'analisi Dell'Economia Agraria
(CREA), Italy

Mary Evelyn Northridge,
New York University, United States

*Correspondence:

Godfred O. Boateng
godfred.boateng@northwestern.edu

Specialty section:

This article was submitted to
Epidemiology,
a section of the journal
Frontiers in Public Health

Received: 26 February 2018

Accepted: 02 May 2018

Published: 11 June 2018

Citation:

Boateng GO, Neilands TB,
Frongillo EA, Melgar-Quinonez HR and
Young SL (2018) Best Practices for
Developing and Validating Scales for
Health, Social, and Behavioral
Research: A Primer.
Front. Public Health 6:149.
doi: 10.3389/fpubh.2018.00149

Scale development and validation are critical to much of the work in the health, social, and behavioral sciences. However, the constellation of techniques required for scale development and evaluation can be onerous, jargon-filled, unfamiliar, and resource-intensive. Further, it is often not a part of graduate training. Therefore, our goal was to concisely review the process of scale development in as straightforward a manner as possible, both to facilitate the development of new, valid, and reliable scales, and to help improve existing ones. To do this, we have created a primer for best practices for scale development in measuring complex phenomena. This is not a systematic review, but rather the amalgamation of technical literature and lessons learned from our experiences spent creating or adapting a number of scales over the past several decades. We identified three phases that span nine steps. In the first phase, items are generated and the validity of their content is assessed. In the second phase, the scale is constructed. Steps in scale construction include pre-testing the questions, administering the survey, reducing the number of items, and understanding how many factors the scale captures. In the third phase, scale evaluation, the number of dimensions is tested, reliability is tested, and validity is assessed. We have also added examples of best practices to each step. In sum, this primer will equip both scientists and practitioners to understand the ontology and methodology of scale development and validation, thereby facilitating the advancement of our understanding of a range of health, social, and behavioral outcomes.

Keywords: scale development, psychometric evaluation, content validity, item reduction, factor analysis, tests of dimensionality, tests of reliability, tests of validity

INTRODUCTION

Scales are a manifestation of latent constructs; they measure behaviors, attitudes, and hypothetical scenarios we expect to exist as a result of our theoretical understanding of the world, but cannot assess directly (1). Scales are typically used to capture a behavior, a feeling, or an action that cannot be captured in a single variable or item. The use of multiple items to measure an underlying latent construct can additionally account for, and isolate, item-specific measurement error, which

leads to more accurate research findings. Thousands of scales have been developed that can measure a range of social, psychological, and health behaviors and experiences.

As science advances and novel research questions are put forth, new scales become necessary. Scale development is not, however, an obvious or a straightforward endeavor. There are many steps to scale development, there is significant jargon within these techniques, the work can be costly and time consuming, and complex statistical analysis is often required. Further, many health and behavioral science degrees do not include training on scale development. Despite the availability of a large amount of technical literature on scale theory and development (1–7), there are a number of incomplete scales used to measure mental, physical, and behavioral attributes that are fundamental to our scientific inquiry (8, 9).

Therefore, our goal is to describe the process for scale development in as straightforward a manner as possible, both to facilitate the development of new, valid, and reliable scales, and to help improve existing ones. To do this, we have created a primer for best practices for scale development. We anticipate this primer will be broadly applicable across many disciplines, especially for health, social, and behavioral sciences. This is not a systematic review, but rather the amalgamation of technical literature and lessons learned from our experiences spent creating or adapting a number of scales related to multiple disciplines (10–23).

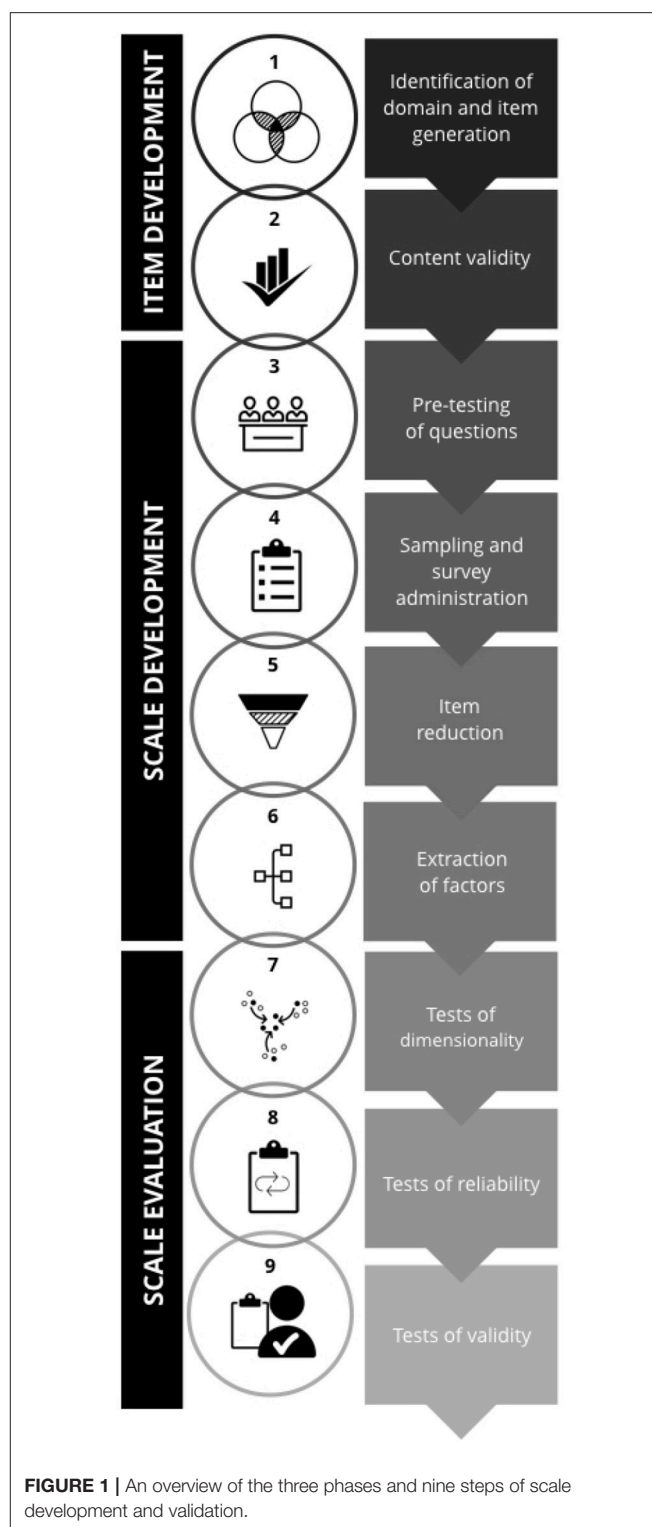
First, we provide an overview of each of the nine steps. Then, within each step, we define key concepts, describe the tasks required to achieve that step, share common pitfalls, and draw on examples in the health, social, and behavioral sciences to recommend best practices. We have tried to keep the material as straightforward as possible; references to the body of technical work have been the foundation of this primer.

SCALE DEVELOPMENT OVERVIEW

There are three phases to creating a rigorous scale—item development, scale development, and scale evaluation (24); these can be further broken down into nine steps (Figure 1).

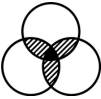


Item development, i.e., coming up with the initial set of questions for an eventual scale, is composed of: (1) identification of the domain(s) and item generation, and (2) consideration of content validity. The second phase, scale development, i.e.,

Abbreviations: A-CASI, audio computer self-assisted interviewing; ASES, adherence self-efficacy scale; CAPI, computer assisted personal interviewing; CFA, confirmatory factor analysis; CASIC, computer assisted survey information collection builder; CFI, comparative fit index; CTT, classical test theory; DIF, differential item functioning; EFA, exploratory factor analysis; FIML, full information maximum likelihood; FNE, fear of negative evaluation; G, global factor; ICC, intraclass correlation coefficient; ICM, Independent cluster model; IRT, item response theory; ODK, Open Data Kit; PAPI, paper and pen/pencil interviewing; QDS, Questionnaire Development System; RMSEA, root mean square error of approximation; SAD, social avoidance and distress; SAS, statistical analysis systems; SASC-R, social anxiety scale for children revised; SEM, structural equation model; SPSS, statistical package for the social sciences; Stata, statistics and data; SRMR, standardized root mean square residual of approximation; TLI, Tucker Lewis Index; WASH, water, sanitation, and hygiene; WRMR, weighted root mean square residual.







turning individual items into a harmonious and measuring construct, consists of (3) pre-testing questions, (4) sampling and survey administration, (5) item reduction, and (6) extraction of latent factors. The last phase, scale evaluation, requires: (7) tests of dimensionality, (8) tests of reliability, and (9) tests of validity.

TABLE 1 | The three phases and nine steps of scale development and validation.

Activity	Purpose	How to explore or estimate?	References
PHASE 1: ITEM DEVELOPMENT			
 Step 1: Identification of Domain and Item Generation: Selecting Which Items to Ask			
Domain identification	To specify the boundaries of the domain and facilitate item generation	1.1 Specify the purpose of the domain 1.2 Confirm that there are no existing instruments 1.3 Describe the domain and provide preliminary conceptual definition 1.4 Specify the dimensions of the domain if they exist <i>a priori</i> 1.5 Define each dimension	(1–4), (25)
Item generation	To identify appropriate questions that fit the identified domain	1.6 Deductive methods: literature review and assessment of existing scales 1.7 Inductive methods: exploratory research methodologies including focus group discussions and interviews	(2–5), (24–41)
 Step 2: Content Validity: Assessing if the Items Adequately Measure the Domain of Interest			
Evaluation by experts	To evaluate each of the items constituting the domain for content relevance, representativeness, and technical quality	2.1 Quantify assessments of 5–7 expert judges using formalized scaling and statistical procedures including content validity ratio, content validity index, or Cohen's coefficient alpha 2.2 Conduct Delphi method with expert judges	(1–5), (24, 42–48)
Evaluation by target population	To evaluate each item constituting the domain for representativeness of actual experience from target population	2.3 Conduct cognitive interviews with end users of scale items to evaluate face validity	(20, 25)
PHASE 2: SCALE DEVELOPMENT			
 Step 3: Pre-testing Questions: Ensuring the Questions and Answers Are Meaningful			
Cognitive interviews	To assess the extent to which questions reflect the domain of interest and that answers produce valid measurements	3.1 Administer draft questions to 5–15 interviewees in 2–3 rounds while allowing respondents to verbalize the mental process entailed in providing answers	(49–54)
 Step 4: Survey Administration and Sample Size: Gathering Enough Data from the Right People			
Survey administration	To collect data with minimum measurement errors	4.1 Administer potential scale items on a sample that reflects range of target population using paper or device	(55–58)
Establishing the sample size	To ensure the availability of sufficient data for scale development	4.2 Recommended sample size is 10 respondents per survey item and/or 200–300 observations	(29, 59–65)
Determining the type of data to use	To ensure the availability of data for scale development and validation	4.3 Use cross-sectional data for exploratory factor analysis 4.4 Use data from a second time point, at least 3 months later in a longitudinal dataset, or an independent sample for test of dimensionality (Step 7)	–
 Step 5: Item Reduction: Ensuring Your Scale Is Parsimonious			
Item difficulty index	To determine the proportion of correct answers given per item (CTT) To determine the probability of a particular examinee correctly answering a given item (IRT)	5.1 Proportion can be calculated for CTT and item difficulty parameter estimated for IRT using statistical packages	(1, 2, 66–68)

(Continued)

TABLE 1 | Continued

Activity	Purpose	How to explore or estimate?	References
Item discrimination test	To determine the degree to which an item or set of test questions are measuring a unitary attribute (CTT) To determine how steeply the probability of correct response changes as ability increases (IRT)	5.2 Estimate biserial correlations or item discrimination parameter using statistical packages	(69–75)
Inter-item and item-total correlations	To determine the correlations between scale items, as well as the correlations between each item and sum score of scale items	5.3 Estimate inter-item/item communalities, item-total, and adjusted item-total correlations using statistical packages	(1, 2, 68, 76)
Distractor efficiency analysis	To determine the distribution of incorrect options and how they contribute to the quality of items	5.4 Estimate distractor analysis using statistical packages	(77–80)
Deleting or imputing missing cases	To ensure the availability of complete cases for scale development	5.5 Delete items with many cases that are permanently missing, or use multiple imputation or full information maximum likelihood for imputation of data	(81–84)
 Step 6: Extraction of Factors: Exploring the Number of Latent Constructs that Fit Your Observed Data			
Factor analysis	To determine the optimal number of factors or domains that fit a set of items	6.1 Use scree plots, exploratory factor analysis, parallel analysis, minimum average partial procedure, and/or the Hull method	(2–4), (85–90)
PHASE 3: SCALE EVALUATION			
 Step 7: Tests of Dimensionality: Testing if Latent Constructs Are as Hypothesized			
Test dimensionality	To address queries on the latent structure of scale items and their underlying relationships. i.e., to validate whether the previous hypothetical structure fits the items	7.1 Estimate independent cluster model—confirmatory factor analysis, cf. Table 2 7.2 Estimate bifactor models to eliminate ambiguity about the type of dimensionality—unidimensionality, bidimensionality, or multi-dimensionality 7.3 Estimate measurement invariance to determine whether hypothesized factor and dimension is congruent across groups or multiple samples	(91–114)
Score scale items	To create scale scores for substantive analysis including reliability and validity of scale	7.4. calculate scale scores using an unweighted approach, which includes summing standardized item scores and raw item scores, or computing the mean for raw item scores 7.5. Calculate scale scores by using a weighted approach, which includes creating factor scores via confirmatory factor analysis or structural equation models	(115)
 Step 8: Tests of Reliability: Establishing if Responses Are Consistent When Repeated			
Calculate reliability statistics	To assess the internal consistency of the scale. i.e., the degree to which the set of items in the scale co-vary, relative to their sum score	8.1 Estimate using Cronbach's alpha 8.2. Other tests such as Raykov's rho, ordinal alpha, and Revelle's beta can be used to assess scale reliability	(116–123)
Test–retest reliability	To assess the degree to which the participant's performance is repeatable; i.e., how consistent their scores are across time	8.3 Estimate the strength of the relationship between scale items over two or three time points; variety of measures possible	(1, 2, 124, 125)
 Step 9: Tests of Validity: Ensuring You Measure the Latent Dimension You Intended			
Criterion validity			
Predictive validity	To determine if scores predict future outcomes	9.1 Use bivariate and multivariable regression; stronger and significant associations or causal effects suggest greater predictive validity	(1, 2, 31)

(Continued)

TABLE 1 | Continued

Activity	Purpose	How to explore or estimate?	References
Concurrent validity	To determine the extent to which scale scores have a stronger relationship with criterion measurements made near the time of administration	9.2 Estimate the association between scale scores and “gold standard” of scale measurement; stronger significant association in Pearson product-moment correlation suggests support for concurrent validity	(2)
Construct validity			
Convergent validity	To examine if the same concept measured in different ways yields similar results	9.3 Estimate the relationship between scale scores and similar constructs using multi-trait multi-method matrix, latent variable modeling, or Pearson product-moment coefficient; higher/stronger correlation coefficients suggest support for convergent validity	(2, 37, 126)
Discriminant validity	To examine if the concept measured is different from some other concept	9.4 Estimate the relationship between scale scores and distinct constructs using multi-trait multi-method matrix, latent variable modeling, or Pearson product-moment coefficient; lower/weaker correlation coefficients suggest support for discriminant validity	(2, 37, 126)
Differentiation by “known groups”	To examine if the concept measured behaves as expected in relation to “known groups”	9.5 Select known binary variables based on theoretical and empirical knowledge and determine the distribution of the scale scores over the known groups; use <i>t</i> -tests if binary, ANOVA if multiple groups	(2, 126)
Correlation analysis	To determine the relationship between existing measures or variables and newly developed scale scores	9.6 Correlate scale scores and existing measures or, preferably, use linear regression, intraclass correlation coefficient, and analysis of standard deviations of the differences between scores	(2, 127, 128)

PHASE 1: ITEM DEVELOPMENT

Step 1: Identification of the Domain(s) and Item Generation

Domain Identification

The first step is to articulate the domain(s) that you are endeavoring to measure. A domain or construct refers to the concept, attribute, or unobserved behavior that is the target of the study (25). Therefore, the domain being examined should be decided upon and defined before any item activity (2). A well-defined domain will provide a working knowledge of the phenomenon under study, specify the boundaries of the domain, and ease the process of item generation and content validation.

McCoach et al. outline a number of steps in scale development; we find the first five to be suitable for the identification of domain (4). These are all based on thorough literature review and include (a) specifying the purpose of the domain or construct you seek to develop, and (b), confirming that there are no existing instruments that will adequately serve the same purpose. Where there is a similar instrument in existence, you need to justify why the development of a new instrument is appropriate and how it will differ from existing instruments. Then, (c) describe the domain and provide a preliminary conceptual definition and (d) specify, if any, the dimensions of the domain. Alternatively, you can let the number of dimensions forming the domain to be determined through statistical computation (cf. Steps 5, 6, and 7). Domains are determined *a priori* if there is an established framework or theory guiding the study, but *a posteriori* if none exist. Finally, if domains are identified *a priori*, (e) the final conceptual definition for each domain should be specified.

Item Generation

Once the domain is delineated, the item pool can then be identified. This process is also called “question development” (26) or “item generation” (24). There are two ways to identify appropriate questions: deductive and inductive methods (24).

The deductive method, also known as “logical partitioning” or “classification from above” (27) is based on the description of the relevant domain and the identification of items. This can be done through literature review and assessment of existing scales and indicators of that domain (2, 24). The inductive method, also known as “grouping” or “classification from below” (24, 27) involves the generation of items from the responses of individuals (24). Qualitative data obtained through direct observations and exploratory research methodologies, such as focus groups and individual interviews, can be used to inductively identify domain items (5).

It is considered best practice to combine both deductive and inductive methods to both define the domain and identify the questions to assess it. While the literature review provides the theoretical basis for defining the domain, the use of qualitative techniques moves the domain from an abstract point to the identification of its manifest forms. A scale or construct defined by theoretical underpinnings is better placed to make specific pragmatic decisions about the domain (28), as the construct will be based on accumulated knowledge of existing items.

It is recommended that the items identified using deductive and inductive approaches should be broader and more comprehensive than one’s own theoretical view of the target (28, 29). Further, content should be included that ultimately will be shown to be tangential or unrelated to the core construct. In other words, one should not hesitate to have items on the

scale that do not perfectly fit the domain identified, as successive evaluation will eliminate undesirable items from the initial pool. Kline and Schinka et al. note that the initial pool of items developed should be at minimum twice as long as the desired final scale (26, 30). Others have recommended the initial pool to be five times as large as the final version, to provide the requisite margin to select an optimum combination of items (30). We agree with Kline and Schinka et al. (26, 30) that the number of items should be at least twice as long as the desired scale.

Further, in the development of items, the *form* of the items, the *wording of the items*, and the types of *responses* that the question is designed to induce should be taken into account. It also means questions should capture the lived experiences of the phenomenon by target population (30). Further, items should be worded simply and unambiguously. Items should not be offensive or potentially biased in terms of social identity, i.e., gender, religion, ethnicity, race, economic status, or sexual orientation (30).

Fowler identified five essential characteristics of items required to ensure the quality of construct measurement (31). These include (a) the need for items to be consistently understood; (b) the need for items to be consistently administered or communicated to respondents; (c) the consistent communication of what constitutes an adequate answer; (d) the need for all respondents to have access to the information needed to answer the question accurately; and (e) the willingness for respondents to provide the correct answers required by the question at all times.

These essentials are sometimes very difficult to achieve. Krosnick (32) suggests that respondents can be less thoughtful about the meaning of a question, search their memories less comprehensively, integrate retrieved information less carefully, or even select a less precise response choice. All this means that they are merely satisficing, i.e., providing merely satisfactory answers, rather than the most accurate ones. In order to combat this behavior, questions should be kept simple, straightforward, and should follow the conventions of normal conversation.

With regards to the type of responses to these questions, we recommend that questions with dichotomous response categories (e.g., true/false) should have no ambiguity. When a Likert-type response scale is used, the points on the scale should reflect the entire measurement continuum. Responses should be presented in an ordinal manner, i.e., in an ascending order without any overlap, and each point on the response scale should be meaningful and interpreted the same way by each participant to ensure data quality (33).

In terms of the number of points on the response scale, Krosnick and Presser (33) showed that responses with just two to three points have lower reliability than Likert-type response scales with five to seven points. However, the gain levels off after seven points. Therefore, response scales with five points are recommended for unipolar items, i.e., those reflecting relative degrees of a single item response quality, e.g., not at all satisfied to very satisfied. Seven response items are recommended for bipolar items, i.e., those reflecting relative degrees of two qualities of an item response scale, e.g., completely dissatisfied to completely satisfied. As an analytic aside, items with scale points fewer

than five categories are best estimated using robust categorical methods. However, items with five to seven categories without strong floor or ceiling effects can be treated as continuous items in confirmatory factor analysis and structural equation modeling using maximum likelihood estimations (34).

One pitfall in the identification of domain and item generation is the improper conceptualization and definition of the domain(s). This can result in scales that may either be deficient because the definition of the domain is ambiguous or has been inadequately defined (35). It can also result in contamination, i.e., the definition of the domain overlaps with other existing constructs in the same field (35).

Caution should also be taken to avoid construct underrepresentation, which is when a scale does not capture important aspects of a construct because its focus is too narrow (35, 36). Further, construct-irrelevant variance, which is the degree to which test scores are influenced by processes that have little to do with the intended construct and seem to be widely inclusive of non-related items (36, 37), should be avoided. Both construct underrepresentation and irrelevant variance can lead to the invalidation of the scale (36).

An example of best practice using the deductive approach to item generation is found in the work of Dennis on breastfeeding self-efficacy (38–40). Dennis' breastfeeding self-efficacy scale items were first informed by Bandura's theory on self-efficacy, followed by content analysis of literature review, and empirical studies on breastfeeding-related confidence.

A valuable example for a rigorous inductive approach is found in the work of Frongillo and Nanama on the development and validation of an experience-based measure of household food insecurity in northern Burkina Faso (41). In order to generate items for the measure, they undertook in-depth interviews with 10 household heads and 26 women using interview guides. The data from these interviews were thematically analyzed, with the results informing the identification of items to be added or deleted from the initial questionnaire. Also, the interviews led to the development and revision of answer choices.

Step 2: Content Validity

Content validity, also known as "theoretical analysis" (5), refers to the "adequacy with which a measure assesses the domain of interest" (24). The need for content adequacy is vital if the items are to measure what they are presumed to measure (1). Additionally, content validity specifies content relevance and content representations, i.e., that the items capture the relevant experience of the target population being examined (129).

Content validity entails the process of ensuring that only the phenomenon spelled out in the conceptual definition, but not other aspects that "might be related but are outside the investigator's intent for that particular [construct] are added" (1). Guion has proposed five conditions that must be satisfied in order for one to claim any form of content validity. We find these conditions to be broadly applicable to scale development in any discipline. These include that (a) the behavioral content has a generally accepted meaning or definition; (b) the domain is unambiguously defined; (c) the content domain is relevant to the purposes of measurement; (d) qualified judges agree that the

domain has been adequately sampled based on consensus; and (e) the response content must be reliably observed and evaluated (42). Therefore, content validity requires evidence of content relevance, representativeness, and technical quality.

Content validity is mainly assessed through evaluation by expert and target population judges.

Evaluation by Experts

Expert judges are highly knowledgeable about the domain of interest and/or scale development; target population judges are potential users of the scale (1, 5). Expert judges seem to be used more often than target-population judges in scale development work to date. Ideally, one should combine expert and target population judgment. When resources are constrained, however, we recommend *at least* the use of expert judges.

Expert judges evaluate each of the items to determine whether they represent the domain of interest. These expert judges should be independent of those who developed the item pool. Expert judgment can be done systematically to avoid bias in the assessment of items. Multiple judges have been used (typically ranging from 5 to 7) (25). Their assessments have been quantified using formalized scaling and statistical procedures such as the content validity ratio for quantifying consensus (43), content validity index for measuring proportional agreement (44), or Cohen's coefficient kappa (k) for measuring inter-rater or expert agreement (45). Among the three procedures, we recommend Cohen's coefficient kappa, which has been found to be most efficient (46). Additionally, an increase in the number of experts has been found to increase the robustness of the ratings (25, 44).

Another way by which content validity can be assessed through expert judges is by using the Delphi method to come to a consensus on which questions are a reflection of the construct you want to measure. The Delphi method is a technique "for structuring group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem" (47).

A good example of evaluation of content validity using expert judges is seen in the work of Augustine et al. on adolescent knowledge of micronutrients (48). After identifying a list of items to be validated, the authors consulted experts in the field of nutrition, psychology, medicine, and basic sciences. The items were then subjected to content analysis using expert judges. Two independent reviews were carried out by a panel of five experts to select the questions that were appropriate, accurate, and interpretable. Items were either accepted, rejected, or modified based on majority opinion (48).

Evaluation by Target Population

Target population judges are experts at evaluating face validity, which is a component of content validity (25). Face validity is the "degree that respondents or end users [or lay persons] judge that the items of an assessment instrument are appropriate to the targeted construct and assessment objectives" (25). These end-users are able to tell whether the construct is a good measure of the domain through cognitive interviews, which we discuss in Step 3.

An example of the concurrent use of expert and target population judges comes from Boateng et al.'s work to develop a household-level water insecurity scale appropriate for use in western Kenya (20). We used the Delphi method to obtain three rounds of feedback from international experts including those in hydrology, geography, WASH and water-related programs, policy implementation, and food insecurity. Each of the three rounds was interspersed with focus group discussions with our target population, i.e., people living in western Kenya. In each round, the questionnaires progressively became more closed ended, until consensus was attained on the definition of the domain we were studying and possible items we could use.

PHASE 2: SCALE DEVELOPMENT

Step 3: Pre-testing Questions

Pre-testing helps to ensure that items are meaningful to the target population before the survey is actually administered, i.e., it minimizes misunderstanding and subsequent measurement error. Because pre-testing eliminates poorly worded items and facilitates revision of phrasing to be maximally understood, it also serves to reduce the cognitive burden on research participants. Finally, pre-testing represents an additional way in which members of the target population can participate in the research process by contributing their insights to the development of the survey.

Pre-testing has two components: the first is the examination of the extent to which the questions reflect the domain being studied. The second is the examination of the extent to which answers to the questions asked produce valid measurements (31).

Cognitive Interviews

To evaluate whether the questions reflect the domain of study and meet the requisite standards, techniques including cognitive interviews, focus group discussion, and field pre-testing under realistic conditions can be used. We describe the most recommended, which is cognitive interviews.

Cognitive interviewing entails the administration of draft survey questions to target populations and then asking the respondents to verbalize the mental process entailed in providing such answers (49). Generally, cognitive interviews allow for questions to be modified, clarified, or augmented to fit the objectives of the study. This approach helps to determine whether the question is generating the information that the author intends by helping to ensure that respondents understand questions as developers intended and that respondents are able to answer in a manner that reflects their experience (49, 50). This can be done on a sample outside of the study population or on a subset of study participants, but it must be explored before the questionnaire is finalized (51, 52).

The sample used for cognitive interviewing should capture the range of demographics you anticipate surveying (49). A range of 5–15 interviews in two to three rounds, or until saturation, or relatively few new insights emerge is considered ideal for pre-testing (49, 51, 52).

In sum, cognitive interviews get to the heart of both assessing the appropriateness of the question to the target population and the strength of the responses (49). The advantages of using cognitive interviewing include: (a) it ensures questions are producing the intended data, (b) questions that are confusing to participants are identified and improved for clarity, (c) problematic questions or questions that are difficult to answer are identified, (d) it ensures response options are appropriate and adequate, (e) it reveals the thought process of participants on domain items, and (f) it can indicate problematic question order (52, 53). Outcomes of cognitive interviews should always be reported, along with solutions used to remedy the situation.

An example of best practice in pre-testing is seen in the work of Morris et al. (54). They developed and validated a novel scale for measuring interpersonal factors underlying injection drug use behaviors among injecting partners. After item development and expert judgment, they conducted cognitive interviews with seven respondents with similar characteristics to the target population to refine and assess item interpretation and to finalize item structure. Eight items were dropped after cognitive interviews for lack of clarity or importance. They also made modifications to grammar, word choice, and answer options based on the feedback from cognitive interviews.

Step 4: Survey Administration and Sample Size

Survey Administration

Collecting data with minimum measurement errors from an adequate sample size is imperative. These data can be collected using paper and pen/pencil interviewing (PAPI) or Computer Assisted Personal Interviewing (CAPI) on devices like laptops, tablets, or phones. A number of software programs exist for building forms on devices. These include Computer Assisted Survey Information Collection (CASIC) Builder™ (West Portal Software Corporation, San Francisco, CA); Qualtrics Research Core™ (www.qualtrics.com); Open Data Kit (ODK, <https://opendatakit.org/>); Research Electronic Data Capture (REDCap) (55); SurveyCTO (Dobility, Inc. <https://www.surveyccto.com/>); and Questionnaire Development System™ (QDS, www.novaresearch.com), which allows the participant to report sensitive audio data.

Each approach has advantages and drawbacks. Using technology can reduce the errors associated with data entry, allow the collection of data from large samples with minimal cost, increase response rate, reduce enumerator errors, permit instant feedback, and increase monitoring of data collection and ability to get more confidential data (56–58, 130). A subset of technology-based programs offers the option of attaching audio files to the survey questions so that questions may be recorded and read out loud to participants with low literacy via audio computer self-assisted interviewing (A-CASI) (131). Self-interviewing, whether via A-CASI or via computer-assisted personal interviewing, in which participants read and respond to questions on a computer without interviewer involvement, may increase reports of sensitive or stigmatized behaviors such

as sexual behaviors and substance use, compared to when being asked by another human.

On the other hand, paper forms may avert the crisis of losing data if the software crashes, the devices are lost or stolen prior to being backed up, and may be more suitable in areas that have irregular electricity and/or internet. However, as sample sizes increase, the use of PAPI becomes more expensive, time and labor intensive, and the data are exposed in several ways to human error (57, 58). Based on the merits of CAPI over PAPI, we recommend researchers use CAPI in data collection for surveys when feasible.

Establishing the Sample Size

The sample size to use for the development of a latent construct has often been contentious. It is recommended that potential scale items be tested on a heterogeneous sample, i.e., a sample that both reflects and captures the range of the target population (29). For example, when the scale is used in a clinical setting, Clark and Watson recommend using patient samples early on instead of a sample from the general population (29).

The necessary sample size is dependent on several aspects of any given study, including the level of variation between the variables, and the level of over-determination (i.e., the ratio of variables to number of factors) of the factors (59). The rule of thumb has been at least 10 participants for each scale item, i.e., an ideal ratio of respondents to items is 10:1 (60). However, others have suggested sample sizes that are independent of the number of survey items. Clark and Watson (29) propose using 300 respondents after initial pre-testing. Others have recommended a range of 200–300 as appropriate for factor analysis (61, 62). Based on their simulation study using different sample sizes, Guadagnoli and Velicer (61) suggested that a minimum of 300–450 is required to observe an acceptable comparability of patterns, and that replication is required if the sample size is <300. Comrey and Lee suggest a graded scale of sample sizes for scale development: 100 = poor, 200 = fair, 300 = good, 500 = very good, ≥1,000 = excellent (63). Additionally, item reduction procedures (described, below in Step 5), such as parallel analysis which requires bootstrapping (estimating statistical parameters from sample by means of resampling with replacement) (64), may require larger data sets.

In sum, there is no single item-ratio that works for all survey development scenarios. A larger sample size or respondent: item ratio is always better, since a larger sample size implies lower measurement errors and more, stable factor loadings, replicable factors, and generalizable results to the true population structure (59, 65). A smaller sample size or respondent: item ratio may mean more unstable loadings and factors, random, non-replicable factors, and non-generalizable results (59, 65). Sample size is, however, always constrained by resources available, and more often than not, scale development can be difficult to fund.

Determining the Type of Data to Use

The development of a scale minimally requires data from a single point in time. To fully test for the reliability of the scale (cf. Steps 8, 9), however, either an independent dataset or a subsequent time point is necessary. Data from longitudinal studies can be

used for initial scale development (e.g., from baseline), to conduct confirmatory factor analysis (using follow-up data, cf. Step 7), and to assess test–retest reliability (using baseline and follow-up data). The problem with using longitudinal data to test hypothesized latent structures is common error variance, since the same, potentially idiosyncratic, participants will be involved. To give the most credence to the reliability of scale, the ideal procedure is to develop the scale on sample A, whether cross-sectional or longitudinal, and then test it on an independent sample B.

The work of Chesney et al. on the Coping Self-Efficacy scale provides an example of this best practice in the use of independent samples (132). This study sought to investigate the psychometric characteristics of the Coping Self-Efficacy (CSE) scale, and their samples came from two independent randomized clinical trials. As such, two independent samples with four different time points each (0, 3, 6, and 12 months) were used. The authors administered the 26-item scale to the sample from the first clinical trial and examined the covariance that existed between all the scale items (exploratory factor analysis) giving the hypothesized factor structure across time in that one trial. The obtained factor structure was then fitted to baseline data from the second randomized clinical trial to test the hypothesized factor structure generated in the first sample (132).

Step 5: Item Reduction Analysis

In scale development, item reduction analysis is conducted to ensure that only parsimonious, functional, and internally consistent items are ultimately included (133). Therefore, the goal of this phase is to identify items that are not or are the least-related to the domain under study for deletion or modification.

Two theories, Classical Test Theory (CTT) and the Item Response Theory (IRT), underpin scale development (134). CTT is considered the traditional test theory and IRT the modern test theory; both function to produce latent constructs. Each theory may be used singly or in conjunction to complement the other's strengths (15, 135). Whether the researcher is using CTT or IRT, the primary goal is to obtain functional items (i.e., items that are correlated with each other, discriminate between individual cases, underscore a single or multidimensional domain, and contribute significantly to the construct).

CTT allows the prediction of outcomes of constructs and the difficulty of items (136). CTT models assume that items forming constructs in their observed, manifest forms consist of a true score on the domain of interest and a random error (which is the differences between the true score and a set of observed scores by an individual) (137). IRT seeks to model the way in which constructs manifest themselves in terms of observable item response (138). Comparatively, the IRT approach to scale development has the advantage of allowing the researcher to determine the effect of adding or deleting a given item or set of items by examining the item information and standard error functions for the item pool (138).

Several techniques exist within the two theories to reduce the item pool, depending on which test theory is driving the scale. The five major techniques used are: item difficulty and item discrimination indices, which are primarily for binary responses;

inter-item and item-total correlations, which are mostly used for categorical items; and distractor efficiency analysis for items with multiple choice response options (1, 2).

Item Difficulty Index

The item difficulty index is both a CTT and an IRT parameter that can be traced largely to educational and psychological testing to assess the relative difficulties and discrimination abilities of test items (66). Subsequently, this approach has been applied to more attitudinal-type scales designed to measure latent constructs.

Under the CTT framework, the item difficulty index, also called item easiness, is the proportion of correct answers on a given item, e.g., the proportion of correct answers on a math test (1, 2). It ranges between 0.0 and 1.0. A high difficulty score means a greater proportion of the sample answered the question correctly. A lower difficulty score means a smaller proportion of the sample understood the question and answered correctly. This may be due to the item being coded wrongly, ambiguity with the item, confusing language, or ambiguity with response options. A lower difficulty score suggests a need to modify the items or delete them from the pool of items.

Under the IRT framework, the item difficulty parameter is the probability of a particular examinee correctly answering any given item (67). This has the advantage of allowing the researcher to identify the different levels of individual performance on specific questions, as well as develop particular questions to specific subgroups or populations (67). Item difficulty is estimated directly using logistic models instead of proportions.

Researchers must determine whether they need items with low, medium, or high difficulty. For instance, researchers interested in general purpose scales will focus on items with medium difficulty (68), i.e., the proportion with item assertions ranging from 0.4 to 0.6 (2, 68). The item difficulty index can be calculated using existing commands in *Mplus*, R, SAS, SPSS, or Stata.

Item Discrimination Index

The item discrimination index (also called item-effectiveness test), is the degree to which an item correctly differentiates between respondents or examinees on a construct of interest (69), and can be assessed under both CTT and IRT frameworks. It is a measure of the difference in performance between groups on a construct. The upper group represents participants with high scores and the lower group those with poor or low scores. The item discrimination index is “calculated by subtracting the proportion of examinees in the lower group (lower %) from the proportion of examinees in the upper group (upper %) who got the item correct or endorsed the item in the expected manner” (69). It differentiates between the number of students in an upper group who get an item correct and the number of students in a lower group who get the item correct (70). The use of an item discrimination index enables the identification of positively discriminating items (i.e., items that differentiate rightly between those who are knowledgeable about a subject and those who are not), negatively discriminating items (i.e., items which are poorly designed such that the more knowledgeable get them wrong and the less knowledgeable get them right), and non-discriminating

item (i.e., items that fail to differentiate between participants who are knowledgeable about a subject and those who are not) (70).

The item discrimination index has been found to improve test items in at least three ways. First, non-discriminating items, which fail to discriminate between respondents because they may be too easy, too hard, or ambiguous, should be removed (71). Second, items which negatively discriminate, e.g., items which fail to differentiate rightly between medically diagnosed depressed and non-depressed respondents on a happiness scale, should be reexamined and modified (70, 71). Third, items that positively discriminate should be retained, e.g., items that are correctly affirmed by a greater proportion of respondents who are medically free of depression, with very low affirmation by respondents diagnosed to be medically depressed (71). In some cases, it has been recommended that such positively discriminating items be considered for revision (70) as the differences could be due to the level of difficulty of the item.

An item discrimination index can be calculated through correlational analysis between the performance on an item and an overall criterion (69) using either the point biserial correlation coefficient or the phi coefficient (72).

Item discrimination under the IRT framework is a slope parameter that determines how steeply the probability of a correct response changes as the proficiency or trait increases (73). This allows differentiation between individuals with similar abilities and can also be estimated using a logistic model. Under certain conditions, the biserial correlation coefficient under the CTT framework has proven to be identical to the IRT item discrimination parameter (67, 74, 75); thus, as the trait increases so does the probability of endorsing an item. These parameters can be computed using existing commands in *Mplus*, R, SAS, SPSS, or Stata. In both CTT and IRT, higher values are indicators of greater discrimination (73).

Inter-item and Item-Total Correlations

A third technique to support the deletion or modification of items is the estimation of inter-item and item-total correlations, which falls under CTT. These correlations often displayed in the form of a matrix are used to examine relationships that exist between individual items in a pool.

Inter-item correlations (also known as polychoric correlations for categorical variables and tetrachoric correlations for binary items) examines the extent to which scores on one item are related to scores on all other items in a scale (2, 68, 76). Also, it examines the extent to which items on a scale are assessing the same content (76). Items with very low correlations (<0.30) are less desirable and could be a cue for potential deletion from the tentative scale.

Item-total correlations (also known as polyserial correlations for categorical variables and biserial correlations for binary items) aim at examining the relationship between each item vs. the total score of scale items. However, the adjusted item-total correlation, which examines the correlation between the item and the sum score of the rest of the items excluding itself is preferred (1, 2). Items with very low adjusted item-total correlations (<0.30) are less desirable and could be a cue for potential deletion from

the tentative scale. Inter-item and item total correlations can be calculated using *Mplus*, R, SAS, SPSS, or Stata.

Distractor Efficiency Analysis

The distractor efficiency analysis shows the distribution of incorrect options and how they contribute to the quality of a multiple-choice item (77). The incorrect options, also known as distractors, are intentionally added in the response options to attract students who do not know the correct answer in a test question (78). To calculate this, respondents will be grouped into three groups—high, middle, and lower tertiles based on their total scores on a set of items. Items will be regarded as appropriate if 100% of those in the high group choose the correct response options, about 50% of those in the middle choose the correct option, and few or none in the lower group choose the correct option (78). This type of analysis is rarely used in the health sciences, as most multiple-choice items are on a Likert-type response scale and do not test respondent correct knowledge, but their experience or perception. However, distractor analysis can help to determine whether items are well-constructed, meaningful, and functional when researchers add response options to questions that do not fit a particular experience. It is expected that participants who are determined as having poor knowledge or experience on the construct will choose the distractors, while those with the right knowledge and experience will choose the correct response options (77, 79). Where those with the right knowledge and experience are not able to differentiate between distractors and the right response, the question may have to be modified. Non-functional distractors identified need to be removed and replaced with efficient distractors (80).

Missing Cases

In addition to these techniques, some researchers opt to delete items with large numbers of cases that are missing, when other missing data-handling techniques cannot be used (81). For cases where modern missing data handling can be used, however, several techniques exist to solve the problem of missing cases. Two of the approaches have proven to be very useful for scale development: full information maximum likelihood (FIML) (82) and multiple imputation (83). Both methods can be applied using existing commands in statistical packages such as *Mplus*, R, SAS, and Stata. When using multiple imputation to recover missing data in the context of survey research, the researcher can impute individual items prior to computing scale scores or impute the scale scores from other scale scores (84). However, item-level imputation has been shown to produce more efficient estimates over scale-level imputation. Thus, imputing individual items before scale development is a preferred approach to imputing newly developed scales for missing cases (84).

Step 6: Extraction of Factors

Factor extraction is the phase in which the optimal number of factors, sometimes called domains, that fit a set of items are determined. This is done using factor analysis. Factor analysis is a regression model in which observed standardized variables are regressed on unobserved (i.e., latent) factors. Because the

variables and factors are standardized, the bivariate regression coefficients are also correlations, representing the loading of each observed variable on each factor. Thus, factor analysis is used to understand the latent (internal) structure of a set of items, and the extent to which the relationships between the items are internally consistent (4). This is done by extracting latent factors which represent the shared variance in responses among the multiple items (4). The emphasis is on the number of factors, the salience of factor loading estimates, and the relative magnitude of residual variances (2).

A number of analytical processes have been used to determine the number of factors to retain from a list of items, and it is beyond the scope of this paper to describe all of them. For scale development, commonly available methods to determine the number of factors to retain include a scree plot (85), the variance explained by the factor model, and the pattern of factor loadings (2). Where feasible, researchers could also assess the optimal number of factors to be drawn from the list of items using either parallel analysis (86), minimum average partial procedure (87), or the Hull method (88, 89).

The extraction of factors can also be used to reduce items. With factor analysis, items with factor loadings or slope coefficients that are below 0.30 are considered inadequate as they contribute <10% variation of the latent construct measured. Hence, it is often recommended to retain items that have factor loadings of 0.40 and above (2, 60). Also, items with cross-loadings or that appear not to load uniquely on individual factors can be deleted. For single-factor models in which Rasch IRT modeling is used, items are selected as having a good fit based on mean-square residual summary statistics (infit and outfit) >0.4 and <1.6 (90).

A number of scales developed stop at this phase and jump to tests of reliability, but the factors extracted at this point only provide a *hypothetical* structure of the scale. The dimensionality of these factors need to be tested (cf. Step 7) before moving on to reliability (cf. Step 8) and validity (cf. Step 9) assessment.

PHASE 3: SCALE EVALUATION

Step 7: Tests of Dimensionality

The test of dimensionality is a test in which the hypothesized factors or factor structure extracted from a previous model is tested at a different time point in a longitudinal study or, ideally, on a new sample (91). Tests of dimensionality determine whether the measurement of items, their factors, and function are the same across two independent samples or within the same sample at different time points. Such tests can be conducted using independent cluster model (ICM)-confirmatory factor analysis, bifactor modeling, or measurement invariance.

Confirmatory Factor Analysis

Confirmatory factor analysis is a form of psychometric assessment that allows for the systematic comparison of an alternative *a priori* factor structure based on systematic fit assessment procedures and estimates the relationship between latent constructs, which have been corrected for measurement errors (92). Morin et al. (92) note that it relies on a highly

restrictive ICM, in which cross-loadings between items and non-target factors are assumed to be exactly zero. The systematic fit assessment procedures are determined by meaningful satisfactory thresholds; **Table 2** contains the most common techniques for testing dimensionality. These techniques include the chi-square test of exact fit, Root Mean Square Error of Approximation ($RMSEA \leq 0.06$), Tucker Lewis Index ($TLI \geq 0.95$), Comparative Fit Index ($CFI \geq 0.95$), Standardized Root Mean Square Residual ($SRMR \leq 0.08$), and Weighted Root Mean Square Residual ($WRMR \leq 1.0$) (90, 92–101).

Bifactor Modeling

Bifactor modeling, also referred to as nested factor modeling, is a form of item response theory used in testing dimensionality of a scale (102, 103). This method can be used when the hypothesized factor structure from the previous model produces partially overlapping dimensions so that one could be seeing most of the items loading onto one factor and a few items loading onto a second and/or a third factor. The bifactor model allows researchers to estimate a unidimensional construct while recognizing the multidimensionality of the construct (104, 105). The bifactor model assumes each item loads onto two dimensions, i.e., items forming the construct may be associated with more than one source of true score variance (92). The first is a general latent factor that underlies all the scale items and the second, a group factor (subscale). A “bifactor model is based on the assumption that a *f*-factor solution exists for a set of *n* items with one [general]/Global (G) factor and *f* – 1 Specific (S) factors also called group factors” (92). This approach allows researchers to examine any distortion that may occur when unidimensional IRT models are fit to multidimensional data (104, 105). To determine whether to retain a construct as unidimensional or multidimensional, the factor loadings from the general factor are then compared to those from the group factors (103, 106). Where the factor loadings on the general factor are significantly larger than the group factors, a unidimensional scale is implied (103, 104). This method is assessed based on meaningful satisfactory thresholds. Alternatively, one can test for the coexistence of a general factor that underlies the construct and multiple group factors that explain the remaining variance not explained by the general factor (92). Each of these methods can be done using statistical software such as *Mplus*, R, SAS, SPSS, or Stata.

Measurement Invariance

Another method to test dimensionality is measurement invariance, also referred to as factorial invariance or measurement equivalence (107). Measurement invariance concerns the extent to which the psychometric properties of the observed indicators are transportable (generalizable) across groups or over time (108). These properties include the hypothesized factor structure, regression slopes, intercept, and residual variances. Measurement invariance is tested sequentially at five levels—configural, metric, scalar, strict (residual), and structural (107, 109). Of key significance to the test of dimensionality is configural invariance, which is concerned with whether the hypothesized factor structure is the same across

TABLE 2 | Description of model fit indices and thresholds for evaluating scales developed for health, social, and behavioral research.

Model fit indices	Description	Recommended threshold to use	References
Chi-square test	The chi-square value is a test statistic of the goodness of fit of a factor model. It compares the observed covariance matrix with a theoretically proposed covariance matrix	Chi-square test of model fit has been assessed to be overly sensitive to sample size and to vary when dealing with non-normal variables. Hence, the use of non-normal data, a small sample size ($n = 180\text{--}300$), and highly correlated items make the chi-square approximation inaccurate. An alternative to this is to use the Satorra-Bentler scaled (mean-adjusted) difference chi-squared statistic. The DIFFTEST has been recommended for models with binary and ordinal variables	(2, 93)
Root Mean Squared Error of Approximation (RMSEA)	RMSEA is a measure of the estimated discrepancy between the population and model-implied population covariance matrices per degree of freedom (139).	Browne and Cudeck recommend $RMSEA \leq 0.05$ as indicative of close fit, $0.05 \leq RMSEA \leq 0.08$ as indicative of fair fit, and values >0.10 as indicative of poor fit between the hypothesized model and the observed data. However, Hu and Bentler have suggested $RMSEA \leq 0.06$ may indicate a good fit	(26, 96–100)
Tucker Lewis Index (TLI)	TLI is based on the idea of comparing the proposed factor model to a model in which no interrelationships at all are assumed among any of the items	Bentler and Bonnett suggest that models with overall fit indices of <0.90 are generally inadequate and can be improved substantially. Hu and Bentler recommend $TLI \geq 0.95$	(95–98)
Comparative Fit Index (CFI)	CFI is an incremental relative fit index that measures the relative improvement in the fit of a researcher's model over that of a baseline model	$CFI \geq 0.95$ is often considered an acceptable fit	(95–98)
Standardized Root Mean Square Residual (SRMR)	SRMR is a measure of the mean absolute correlation residual, the overall difference between the observed and predicted correlations	Threshold for acceptable model fit is $SRMR \leq 0.08$	(95–98)
Weighted Root Mean Square Residual (WRMR)	WRMR uses a "variance-weighted approach especially suited for models whose variables measured on different scales or have widely unequal variances" (139); it has been assessed to be most suitable in assessing models fitted to binary and ordinal data	Yu recommends a threshold of $WRMR < 1.0$ for assessing model fit. This index is used for confirmatory factor analysis and structural equation models with binary and ordinal variables	(101)
Standard of Reliability for scales	A reliability of 0.90 is the minimum recommended threshold that should be tolerated while a reliability of 0.95 should be the desirable standard. While the ideal has rarely been attained by most researchers, a reliability coefficient of 0.70 has often been accepted as satisfactory for most scales	Nunnally recommends a threshold of ≥ 0.90 for assessing internal consistency for scales	(117, 123)

groups. This assumption has to be met in order for subsequent tests to be meaningful (107, 109). For example, a hypothesized unidimensional structure, when tested across multiple countries, should be the same. This can be tested in CTT, using multigroup confirmatory factor analysis (110–112).

An alternative approach to measurement invariance in the testing of unidimensionality under item response theory is the Rasch measurement model for binary items and polytomous IRT models for categorical items. Here, emphasis is on testing the differential item functioning (DIF)—an indicator of whether “a group of respondents is scoring better than another group of respondents on an item or a test after adjusting for the overall ability scores of the respondents” (108, 113). This is analogous to the conditions underpinning measurement invariance in a multi-group CFA (108, 113).

Whether the hypothesized structure is bidimensional or multidimensional, each dimension in the structure needs to be tested again to confirm its unidimensionality. This can also be done using confirmatory factor analysis. Appropriate model fit

indices and the strength of factor loadings (cf. **Table 2**) are the basis on which the latent structure of the items can be judged.

One commonly encountered pitfall is a lack of satisfactory global model fit in confirmatory factor analysis conducted on a new sample following a satisfactory initial factor analysis performed on a previous sample. Lack of satisfactory fit offers the opportunity to identify additional underperforming items for removal. Items with very poor loadings (≤ 0.3) can be considered for removal. Also, modification indices, produced by *Mplus* and other structural equation modeling (SEM) programs, can help identify items that need to be modified. Sometimes a higher-order factor structure, where correlations among the original factors can be explained by one or more higher-order factors, is needed. This can also be assessed using statistical software such as *Mplus*, R, SAS, SPSS, or Stata.

A good example of best practice is seen in the work of Pushpanathan et al. on the appropriateness of using a traditional

confirmatory factor analysis or a bifactor model (114) in assessing whether the Parkinson's Disease Sleep Scale-Revised was better used as a unidimensional scale, a tri-dimensional scale, or a scale that has an underlying general factor and three group factors (sub-scales). They tested this using three different models—a unidimensional model (1-factor CFA); a 3-factor model (3 factor CFA) consisting of sub-scales measuring insomnia, motor symptoms and obstructive sleep apnea, and REM sleep behavior disorder; and a confirmatory bifactor model having a general factor and the same three sub-scales combined. The results of this study suggested that only the bifactor model with a general factor and the three sub-scales combined achieved satisfactory model fitness. Based on these results, the authors cautioned against the use of a unidimensional total scale scores as a cardinal indicator of sleep in Parkinson's disease, but encouraged the examination of its multidimensional subscales (114).

Scoring Scale Items

Finalized items from the tests of dimensionality can be used to create scale scores for substantive analysis including tests of reliability and validity. Scale scores can be calculated by using unweighted or weighted procedures. The unweighted approach involves summing standardized item scores or raw item scores, or computing the mean for raw item scores (115). The weighted approach in calculating scale scores can be produced via statistical software programs such as *Mplus*, R, SAS, SPSS, or Stata. For instance, in using confirmatory factor analysis, structural equation models, or exploratory factor analysis, each factor produced reveals a statistically independent source of variation among a set of items (115). The contribution of each individual item to this factor is considered a weight, with the factor loading value representing the weight. The scores associated with each factor in a model then represents a composite scale score based on a weighted sum of the individual items using factor loadings (115). In general, it does not make much difference in the performance of the scale if scales are computed as unweighted items (e.g., mean or sum scores) or weighted items (e.g., factor scores).

Step 8: Tests of Reliability

Reliability is the degree of consistency exhibited when a measurement is repeated under identical conditions (116). A number of standard statistics have been developed to assess reliability of a scale, including Cronbach's alpha (117), ordinal alpha (118, 119) specific to binary and ordinal scale items, test-retest reliability (coefficient of stability) (1, 2), McDonald's Omega (120), Raykov's rho (2) or Revelle's beta (121, 122), split-half estimates, Spearman-Brown formula, alternate form method (coefficient of equivalence), and inter-observer reliability (1, 2). Of these statistics, Cronbach's alpha and test-retest reliability are predominantly used to assess reliability of scales (2, 117).

Cronbach's Alpha

Cronbach's alpha assesses the internal consistency of the scale items, i.e., the degree to which the set of items in the scale co-vary, relative to their sum score (1, 2, 117). An alpha coefficient of 0.70 has often been regarded as an acceptable threshold for reliability;

however, 0.80 and 0.95 is preferred for the psychometric quality of scales (60, 117, 123). Cronbach's alpha has been the most common and seems to have received general approval; however, reliability statistics such as Raykov's rho, ordinal alpha, and Revelle's beta, which are debated to have improvements over Cronbach's alpha, are beginning to gain acceptance.

Test-Retest Reliability

An additional approach in testing reliability is the test-retest reliability. The test-retest reliability, also known as the coefficient of stability, is used to assess the degree to which the participants' performance is repeatable, i.e., how consistent their sum scores are across time (2). Researchers vary in how they assess test-retest reliability. While some prefer to use intra class correlation coefficient (124), others use the Pearson product-moment correlation (125). In both cases, the higher the correlation, the higher the test-retest reliability, with values close to zero indicating low reliability. In addition, study conditions could change values on the construct being measured over time (as in an intervention study, for example), which could lower the test-retest reliability.

The work of Johnson et al. (16) on the validation of the HIV Treatment Adherence Self-Efficacy Scale (ASES) is a good example of the test of reliability. As part of testing for reliability, the authors tested for the internal consistency reliability values for the ASES and its subscales using Raykov's rho (produces a coefficient similar to alpha but with fewer assumptions and with confidence intervals); they then tested for the temporal consistency of the ASES' factor structure. This was then followed by test-retest reliability assessment among the latent factors. The different approaches provided support for the reliability of the ASES scale.

Other approaches found to be useful and support scale reliability include split-half estimates, Spearman-Brown formula, alternate form method (coefficient of equivalence), and inter-observer reliability (1, 2).

Step 9: Tests of Validity

Scale validity is the extent to which "an instrument indeed measures the latent dimension or construct it was developed to evaluate" (2). Although it is discussed at length here in Step 9, validation is an ongoing process that starts with the identification and definition of the domain of study (Step 1) and continues to its generalizability with other constructs (Step 9) (36). The validity of an instrument can be examined in numerous ways; the most common tests of validity are content validity (described in Step 2), which can be done prior to the instrument being administered to the target population, and criterion (predictive and concurrent) and construct validity (convergent, discriminant, differentiation by known groups, correlations), which occurs after survey administration.

Criterion Validity

Criterion validity is the "degree to which there is a relationship between a given test score and performance on another measure of particular relevance, typically referred to as criterion" (1, 2). There are two forms of criterion validity: predictive (criterion)

validity and concurrent (criterion) validity. Predictive validity is “the extent to which a measure predicts the answers to some other question or a result to which it ought to be related with” (31). Thus, the scale should be able to predict a behavior in the future. An example is the ability for an exclusive breastfeeding social support scale to predict exclusive breastfeeding (10). Here, the mother’s willingness to exclusively breastfeed occurs after social support has been given, i.e., it should predict the behavior. Predictive validity can be estimated by examining the association between the scale scores and the criterion in question.

Concurrent criterion validity is the extent to which test scores have a stronger relationship with criterion (gold standard) measurement made at the time of test administration or shortly afterward (2). This can be estimated using Pearson product-moment correlation or latent variable modeling. The work of Greca and Stone on the psychometric evaluation of the revised version of a social anxiety scale for children (SASC-R) provides a good example for the evaluation of concurrent validity (140). In this study, the authors collected data on an earlier validated version of the SASC scale consisting of 10 items, as well as the revised version, SASC-R, which had additional 16 items making a 26-item scale. The SASC consisted of two sub scales [fear of negative evaluation (FNE), social avoidance and distress (SAD)] and the SASC-R produced three new subscales (FNE, SAD-New, and SAD-General). Using a Pearson product-moment correlation, the authors examined the inter-correlations between the common subscales for FNE, and between SAD and SAD-New. With a validity coefficient of 0.94 and 0.88, respectively, the authors found evidence of concurrent validity.

A limitation of concurrent validity is that this strategy for validity does not work with small sample sizes because of their large sampling errors. Secondly, appropriate criterion variables or “gold standards” may not be available (2). This reason may account for its omission in most validation studies.

Construct Validity

Construct validity is the “extent to which an instrument assesses a construct of concern and is associated with evidence that measures other constructs in that domain and measures specific real-world criteria” (2). Four indicators of construct validity are relevant to scale development: convergent validity, discriminant validity, differentiation by known groups, and correlation analysis.

Convergent validity is the extent to which a construct measured in different ways yields similar results. Specifically, it is the “degree to which scores on a studied instrument are related to measures of other constructs that can be expected on theoretical grounds to be close to the one tapped into by this instrument” (2, 37, 126). This is best estimated through the multi-trait multi-method matrix (2), although in some cases researchers have used either latent variable modeling or Pearson product-moment correlation based on Fisher’s Z transformation. Evidence of convergent validity of a construct can be provided by the extent to which the newly developed scale correlates highly with other variables designed to measure the same construct (2, 126). It can be invalidated by too low or weak correlations

with other tests which are intended to measure the same construct.

Discriminant validity is the extent to which a measure is novel and not simply a reflection of some other construct (126). Specifically, it is the “degree to which scores on a studied instrument are differentiated from behavioral manifestations of other constructs, which on theoretical grounds can be expected not to be related to the construct underlying the instrument under investigation” (2). This is best estimated through the multi-trait multi method matrix (2). Discriminant validity is indicated by predictably low or weak correlations between the measure of interest and other measures that are supposedly not measuring the same variable or concept (126). The newly developed construct can be invalidated by too high correlations with other tests which are intended to differ in their measurements (37). This approach is critical in differentiating the newly developed construct from other rival alternatives (36).

Differentiation or comparison between known groups examines the distribution of a newly developed scale score over known binary items (126). This is premised on previous theoretical and empirical knowledge of the performance of the binary groups. An example of best practice is seen in the work of Boateng et al. on the validation of a household water insecurity scale in Kenya. In this study, we compared the mean household water insecurity scores over households with or without *E. coli* present in their drinking water. Consistent with what we knew from the extant literature, we found households with *E. coli* present in their drinking water had higher mean water insecurity scores than households that had no *E. coli* in drinking water. This suggested our scale could discriminate between particular known groups.

Although correlational analysis is frequently used by several scholars, bivariate regression analysis is preferred to correlational analysis for quantifying validity (127, 128). Regression analysis between scale scores and an indicator of the domain examined has a number of important advantages over correlational analysis. First, regression analysis quantifies the association in meaningful units, facilitating judgment of validity. Second, regression analysis avoids confounding validity with the underlying variation in the sample and therefore the results from one sample are more applicable to other samples in which the underlying variation may differ. Third, regression analysis is preferred because the regression model can be used to examine discriminant validity by adding potential alternative measures. In addition to regression analysis, alternative techniques such as analysis of standard deviations of the differences between scores and the examination of intraclass correlation coefficients (ICC) have been recommended as viable options (128).

Taken together, these methods make it possible to assess the validity of an adapted or a newly developed scale. In addition to predictive validity, existing studies in fields such as health, social, and behavioral sciences have shown that scale validity is supported if at least two of the different forms of construct validity discussed in this section have been examined. Further information about establishing validity and constructing indicators from scales can be found in Frongillo et al. (141).

CONCLUSIONS

In sum, we have sought to give an overview of the key steps in scale development and validation (**Figure 1**) as well as to help the reader understand how one might approach each step (**Table 1**). We have also given a basic introduction to the conceptual and methodological underpinnings of each step.

Because scale development is so complicated, this should be considered a primer, i.e., a “jumping off point” for anyone interested in scale development. The technical literature and examples of rigorous scale development mentioned throughout will be important for readers to pursue. There are a number of matters not addressed here, including how to interpret scale output, the designation of cut-offs, when indices, rather than scales, are more appropriate, and principles for re-testing scales in new populations. Also, this review leans more toward the classical test theory approach to scale development; a comprehensive review on IRT modeling will be complementary. We hope this review helps to ease readers into the literature, but space precludes consideration of all these topics.

The necessity of the nine steps that we have outlined here (**Table 1**, **Figure 1**) will vary from study to study. While studies focusing on developing scales *de novo* may use all nine steps, others, e.g., those that set out to validate existing scales, may end up using only the last four steps. Resource constraints, including time, money, and participant attention and patience are very real, and must be acknowledged as additional limits to rigorous scale development. We cannot state which steps are the most important; difficult decisions about which steps to approach less rigorously can only be made by each scale developer, based on the purpose of the research, the proposed end-users of the scale, and resources available. It is our hope, however, that by outlining the general shape of the phases and steps in scale development, researchers will be able to purposively choose the steps that

they will include, rather than omitting a step out of lack of knowledge.

Well-designed scales are the foundation of much of our understanding of a range of phenomena, but ensuring that we accurately quantify what we purport to measure is not a simple matter. By making scale development more approachable and transparent, we hope to facilitate the advancement of our understanding of a range of health, social, and behavioral outcomes.

AUTHOR CONTRIBUTIONS

GB and SY developed the first draft of the scale development and validation manuscript. All authors participated in the editing and critical revision of the manuscript and approved the final version of the manuscript for publication.

FUNDING

Funding for this work was obtained by SY through the National Institute of Mental Health—R21 MH108444. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

ACKNOWLEDGMENTS

We would like to acknowledge the importance of the works of several scholars of scale development and validation used in developing this primer, particularly Robert DeVellis, Tenko Raykov, George Marcoulides, David Streiner, and Betsy McCoach. We would also like to acknowledge the help of Josh Miller of Northwestern University for assisting with design of **Figure 1** and development of **Table 1**, and we thank Zeina Jamuladdine for helpful comments on tests of unidimensionality.

REFERENCES

- DeVellis RF. *Scale Development: Theory and Application*. Los Angeles, CA: Sage Publications (2012).
- Raykov T, Marcoulides GA. *Introduction to Psychometric Theory*. New York, NY: Routledge, Taylor & Francis Group (2011).
- Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford University Press (2015).
- McCoach DB, Gable RK, Madura, JP. *Instrument Development in the Affective Domain. School and Corporate Applications, 3rd Edn*. New York, NY: Springer (2013).
- Morgado FFR, Meireles JFF, Neves CM, Amaral ACS, Ferreira MEC. Scale development: ten main limitations and recommendations to improve future research practices. *Psicol Reflex E Crítica* (2018) **30**:3. doi: 10.1186/s41155-016-0057-1
- Glanz K, Rimer BK, Viswanath K. *Health Behavior: Theory, Research, and Practice*. San Francisco, CA: John Wiley & Sons, Inc (2015).
- Ajzen I. From intentions to actions: a theory of planned behavior. In: *Action Control SSSP Springer Series in Social Psychology* Berlin; Heidelberg: Springer, (1985). p. 11–39.
- Bai Y, Peng C-YJ, Fly AD. Validation of a short questionnaire to assess mothers' perception of workplace breastfeeding support. *J Acad Nutr Diet* (2008) **108**:1221–5. doi: 10.1016/j.jada.2008.04.018
- Hirani SAA, Karmaliani R, Christie T, Rafique G. Perceived Breastfeeding Support Assessment Tool (PBSAT): development and testing of psychometric properties with Pakistani urban working mothers. *Midwifery* (2013) **29**:599–607. doi: 10.1016/j.midw.2012.05.003
- Boateng GO, Martin S., Collins S, Natamba BK, Young SL. Measuring exclusive breastfeeding social support: scale development and validation in Uganda. *Matern Child Nutr.* (2018). doi: 10.1111/mcn.12579. [Epub ahead of print].
- Arbach A, Natamba BK, Achan J, Griffiths JK, Stoltzfus RJ, Mehta S, et al. Reliability and validity of the center for epidemiologic studies-depression scale in screening for depression among HIV-infected and -uninfected pregnant women attending antenatal services in northern Uganda: a cross-sectional study. *BMC Psychiatry* (2014) **14**:303. doi: 10.1186/s12888-014-0303-y
- Natamba BK, Kilama H, Arbach A, Achan J, Griffiths JK, Young SL. Reliability and validity of an individually focused food insecurity access scale for assessing inadequate access to food among pregnant Ugandan women of mixed HIV status. *Public Health Nutr.* (2015) **18**:2895–905. doi: 10.1017/S1368980014001669
- Neilands TB, Chakravarty D, Darbes LA, Beougher SC, Hoff CC. Development and validation of the sexual agreement investment scale. *J Sex Res.* (2010) **47**:24–37. doi: 10.1080/00224490902916017

14. Neilands TB, Choi K-H. A validation and reduced form of the female condom attitudes scale. *AIDS Educ Prev.* (2002) **14**:158–71. doi: 10.1521/aeap.14.2.158.23903
15. Lippman SA, Neilands TB, Leslie HH, Maman S, MacPhail C, Twine R, et al. Development, validation, and performance of a scale to measure community mobilization. *Soc Sci Med.* (2016) **157**:127–37. doi: 10.1016/j.socscimed.2016.04.002
16. Johnson MO, Neilands TB, Dilworth SE, Morin SF, Remien RH, Chesney MA. The role of self-efficacy in HIV treatment adherence: validation of the HIV treatment adherence self-efficacy scale (HIV-ASES). *J Behav Med.* (2007) **30**:359–70. doi: 10.1007/s10865-007-9118-3
17. Sexton JB, Helmreich RL, Neilands TB, Rowan K, Vella K, Boyden J, et al. The Safety Attitudes Questionnaire: psychometric properties, benchmarking data, and emerging research. *BMC Health Serv Res.* (2006) **6**:44. doi: 10.1186/1472-6963-6-44
18. Wolfe WS, Frongillo EA. Building household food-security measurement tools from the ground up. *Food Nutr Bull.* (2001) **22**:5–12. doi: 10.1177/156482650102200102
19. González W, Jiménez A, Madrigal G, Muñoz LM, Frongillo EA. Development and validation of measure of household food insecurity in urban costa rica confirms proposed generic questionnaire. *J Nutr.* (2008) **138**:587–92. doi: 10.1093/jn/138.3.587
20. Boateng GO, Collins SM, Mbullo P, Wekesa P, Onono M, Neilands T, et al. A novel household water insecurity scale: procedures and psychometric analysis among postpartum women in western Kenya. *PLoS ONE.* (2018). doi: 10.1371/journal.pone.0198591
21. Melgar-Quinonez H, Hackett M. Measuring household food security: the global experience. *Rev Nutr.* (2008) **21**:27s–37s. doi: 10.1590/S1415-52732008000700004
22. Melgar-Quinonez H, Zubietta AC, Valdez E, Whitelaw B, Kaiser L. Validación de un instrumento para vigilar la inseguridad alimentaria en la Sierra de Manantlán, Jalisco. *Salud Pública México* (2005) **47**:413–22. doi: 10.1590/S0036-36342005000600005
23. Hackett M, Melgar-Quinonez H, Uribe MCA. Internal validity of a household food security scale is consistent among diverse populations participating in a food supplement program in Colombia. *BMC Public Health* (2008) **8**:175. doi: 10.1186/1471-2458-8-175
24. Hinkin TR. A review of scale development practices in the study of organizations. *J Manag.* (1995) **21**:967–88. doi: 10.1016/0149-2063(95)90050-0
25. Haynes SN, Richard DCS, Kubany ES. Content validity in psychological assessment: a functional approach to concepts and methods. *Psychol Assess.* (1995) **7**:238–47. doi: 10.1037/1040-3590.7.3.238
26. Kline P. *A Handbook of Psychological Testing. 2nd Edn.* London: Routledge; Taylor & Francis Group (1993).
27. Hunt SD. *Modern Marketing Theory.* Cincinnati: South-Western Publishing (1991).
28. Loevinger J. Objective tests as instruments of psychological theory. *Psychol Rep.* (1957) **3**:635–94. doi: 10.2466/pr0.1957.3.3.635
29. Clarke LA, Watson D. Constructing validity: basic issues in objective scale development. *Psychol Assess.* (1995) **7**:309–19. doi: 10.1037/1040-3590.7.3.309
30. Schinka JA, Velicer WF, Weiner IR. *Handbook of Psychology, Vol. 2, Research Methods in Psychology.* Hoboken, NJ: John Wiley & Sons, Inc. (2012).
31. Fowler FJ. *Improving Survey Questions: Design and Evaluation.* Thousand Oaks, CA: Sage Publications (1995).
32. Krosnick JA. Questionnaire design. In: Vannette DL, Krosnick JA, editors. *The Palgrave Handbook of Survey Research.* Cham: Palgrave Macmillan (2018), pp. 439–55.
33. Krosnick JA, Presser S. Question and questionnaire design. In: Wright JD, Marsden PV, editors. *Handbook of Survey Research.* San Diego, CA: Elsevier (2009), pp. 263–314.
34. Rhemtulla M, Brosseau-Liard PÉ, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychol Methods* (2012) **17**:354–73. doi: 10.1037/a0029315
35. MacKenzie SB, Podsakoff PM, Podsakoff NP. Construct measurement and validation procedures in MIS and behavioral research: integrating new and existing techniques. *MIS Q.* (2011) **35**:293. doi: 10.2307/23044045
36. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *Am Psychol.* (1995) **50**:741–9. doi: 10.1037/0003-066X.50.9.741
37. Campbell DT, Fiske DW. Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychol Bull.* (1959) **56**:81–105. doi: 10.1037/h0046016
38. Dennis C. Theoretical underpinnings of breastfeeding confidence: a self-efficacy framework. *J Hum Lact.* (1999) **15**:195–201. doi: 10.1177/089033449901500303
39. Dennis C-L, Faux S. Development and psychometric testing of the Breastfeeding Self-Efficacy Scale. *Res Nurs Health* (1999) **22**:399–409. doi: 10.1002/(SICI)1098-240X(199910)22:5<399::AID-NUR6>3.0.CO;2-4
40. Dennis C-L. The breastfeeding self-efficacy scale: psychometric assessment of the short form. *J Obstet Gynecol Neonatal Nurs.* (2003) **32**:734–44. doi: 10.1177/0884217503258459
41. Frongillo EA, Nanama S. Development and validation of an experience-based measure of household food insecurity within and across seasons in Northern Burkina Faso. *J Nutr.* (2006) **136**:1409S–19S. doi: 10.1093/jn/136.5.1409S
42. Guion R. Content validity - the source of my discontent. *Appl Psychol Meas.* (1977) **1**:1–10. doi: 10.1177/014662167700100103
43. Lawshe C. A quantitative approach to content validity. *Pers Psychol.* (1975) **28**:563–75. doi: 10.1111/j.1744-6570.1975.tb01393.x
44. Lynn M. Determination and quantification of content validity. *Nurs Res.* (1986) **35**:382–5. doi: 10.1097/00006199-198611000-00017
45. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* (1960) **20**:37–46. doi: 10.1177/001316446002000104
46. Wynd CA, Schmidt B, Schaefer MA. Two quantitative approaches for estimating content validity. *West J Nurs Res.* (2003) **25**:508–18. doi: 10.1177/0193945903252998
47. Linstone HA, Turoff M. (eds). *The Delphi Method.* Reading, MA: Addison-Wesley (1975).
48. Augustine LF, Vazir S, Rao SF, Rao MV, Laxmaiah A, Ravinder P, et al. Psychometric validation of a knowledge questionnaire on micronutrients among adolescents and its relationship to micronutrient status of 15–19-year-old adolescent boys, Hyderabad, India. *Public Health Nutr.* (2012) **15**:1182–9. doi: 10.1017/S1368980012000055
49. Beatty PC, Willis GB. Research synthesis: the practice of cognitive interviewing. *Public Opin Q.* (2007) **71**:287–311. doi: 10.1093/poq/nfm006
50. Alaimo K, Olson CM, Frongillo EA. Importance of cognitive testing for survey items: an example from food security questionnaires. *J Nutr Educ.* (1999) **31**:269–75. doi: 10.1016/S0022-3182(99)70463-2
51. Willis GB. *Cognitive Interviewing and Questionnaire Design: A Training Manual. Cognitive Methods Staff Working Paper Series.* Hyattsville, MD: National Center for Health Statistics (1994).
52. Willis GB. *Cognitive Interviewing: A Tool for Improving Questionnaire Design.* Thousand Oaks, CA: Sage Publications (2005).
53. Tourangeau R. Cognitive aspects of survey measurement and mismeasurement. *Int J Public Opin Res.* (2003) **15**:3–7. doi: 10.1093/ijpor/15.1.3
54. Morris MD, Neilands TB, Andrew E, Mahar L, Page KA, Hahn JA. Development and validation of a novel scale for measuring interpersonal factors underlying injection drug using behaviours among injecting partnerships. *Int J Drug Policy* (2017) **48**:54–62. doi: 10.1016/j.drugpo.2017.05.030
55. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* (2009) **42**:377–81. doi: 10.1016/j.jbi.2008.08.010
56. Goldstein M, Benerjee R, Kilic T. *Paper v Plastic Part 1: The Survey Revolution Is in Progress.* The World Bank Development Impact. (2012). Available online at: <http://blogs.worldbank.org/impactevaluations/paper-v-plastic-part-i-the-survey-revolution-is-in-progress> (Accessed November 10, 2017).

57. Fanning J, McAuley E. A Comparison of tablet computer and paper-based questionnaires in healthy aging research. *JMIR Res Protoc.* (2014) 3:e38. doi: 10.2196/resprot.3291
58. Greenlaw C, Brown-Welty S. A Comparison of web-based and paper-based survey methods: testing assumptions of survey mode and response cost. *Eval Rev.* (2009) 33:464–80. doi: 10.1177/0193841X09340214
59. MacCallum RC, Widaman KF, Zhang S, Hong S. Sample size in factor analysis. *Psychol Methods* (1999) 4:84–99. doi: 10.1037/1082-989X.4.1.84
60. Nunnally JC. *Psychometric Theory*. New York, NY: McGraw-Hill (1978).
61. Guadagnoli E, Velicer WF. Relation of sample size to the stability of component patterns. *Am Psychol Assoc.* (1988) 103:265–75. doi: 10.1037/0033-2909.103.2.265
62. Comrey AL. Factor-analytic methods of scale development in personality and clinical psychology. *Am Psychol Assoc.* (1988) 56:754–61.
63. Comrey AL, Lee H. *A First Course in Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. (1992).
64. Ong DC. *A Primer to Bootstrapping and an Overview of doBootstrap*. Stanford, CA: Department of Psychology, Stanford University (2014).
65. Osborne JW, Costello AB. Sample size and subject to item ratio in principal components analysis. *Pract Assess Res Eval.* (2004) 99:1–15. Available online at: <http://pareonline.net/htm/v9n11.htm>
66. Ebel R, Frisbie D. *Essentials of Educational Measurement*. Englewood Cliffs, NJ: Prentice-Hall (1979).
67. Hambleton R, Jones R. An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educ Meas Issues Pract.* (1993) 12:38–47. doi: 10.1111/j.1745-3992.1993.tb00543.x
68. Raykov T. *Scale Construction and Development. Lecture Notes. Measurement and Quantitative Methods*. East Lansing, MI: Michigan State University (2015).
69. Whiston SC. *Principles and Applications of Assessment in Counseling*. Cengage Learning (2008).
70. Brennan RL. A generalized upper-lower item discrimination index. *Educ Psychol Meas.* (1972) 32:289–303. doi: 10.1177/001316447203200206
71. Popham WJ, Husek TR. Implications of criterion-referenced measurement. *J Educ Meas.* (1969) 6:1–9. doi: 10.1111/j.1745-3984.1969.tb00654.x
72. Rasiah S-MS, Isaiah R. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Ann Acad Med Singap.* (2006) 35:67–71. Available online at: <http://repository.um.edu.my/id/eprint/65455>
73. Demars C. *Item Respons Theory*. New York, NY: Oxford University Press (2010).
74. Lord FM. *Applications of Item Response Theory to Practical Testing Problems*. New Jersey, NJ: Englewood Cliffs (1980).
75. Bazaldua DAL, Lee Y-S, Keller B, Fellers L. Assessing the performance of classical test theory item discrimination estimators in Monte Carlo simulations. *Asia Pac Educ Rev.* (2017) 18:585–98. doi: 10.1007/s12564-017-9507-4
76. Piedmont RL. Inter-item correlations. In *Encyclopedia of Quality of Life and Well-Being Research*. Dordrecht: Springer (2014). p. 3303–4. doi: 10.1007/978-94-007-0753-5_1493
77. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med Educ.* (2009) 9:40. doi: 10.1186/1472-6920-9-40
78. Fulcher G, Davidson F. *The Routledge Handbook of Language Testing*. New York, NY: Routledge (2012).
79. Cizek GJ, O'Day DM. Further investigation of nonfunctioning options in multiple-choice test items. *Educ Psychol Meas.* (1994) 54:861–72.
80. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. *Appl Meas Educ.* (1989) 2:51–78. doi: 10.1207/s15324818ame0201_4
81. Tappen RM. *Advanced Nursing Research*. Sudbury, MA: Jones & Bartlett Publishers (2011).
82. Enders CK, Bandalos DL. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Struct Equ Model.* (2009) 8:430–57. doi: 10.1207/S15328007SEM0803_5
83. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Stat Methods Med Res.* (2007) 16:199–218. doi: 10.1177/0962280206075304
84. Gottschall AC, West SG, Enders CK. A Comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivar Behav Res.* (2012) 47:1–25. doi: 10.1080/00273171.2012.640589
85. Cattell RB. The Scree test for the number of factors. *Multivar Behav Res.* (1966) 1: 245–76. doi: 10.1207/s15327906mbr0102_10
86. Horn JL. A rationale and test for the number of factors in factor analysis. *Psychometrika* (1965) 30:179–85. doi: 10.1007/BF02289447
87. Velicer WF. Determining the number of components from the matrix of partial correlations. *Psychometrika* (1976) 41:321–7. doi: 10.1007/BF02293557
88. Lorenzo-Seva U, Timmerman ME, Kiers HAL. The hull method for selecting the number of common factors. *Multivar Behav Res.* (2011) 46:340–64. doi: 10.1080/00273171.2011.564527
89. Jolijn Hendriks AA, Perugini M, Angleitner A, Ostendorf F, Johnson JA, De Fruyt F, et al. The five-factor personality inventory: cross-cultural generalizability across 13 countries. *Eur J Pers.* (2003) 17:347–73. doi: 10.1002/per.491
90. Bond TG, Fox C. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Erlbaum (2013).
91. Brown T. *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford Press (2014).
92. Morin AJS, Arens AK, Marsh HW. A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Struct Equ Model Multidiscip J.* (2016) 23:116–39. doi: 10.1080/10705511.2014.961800
93. Cochran WG. The χ^2 test of goodness of fit. *Ann Math Stat.* (1952) 23:315–45. doi: 10.1214/aoms/1177729380
94. Brown MW. *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford Press (2014).
95. Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* (1973) 38:1–10. doi: 10.1007/BF02291170
96. Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. *Psychol Bull.* (1980) 88:588–606. doi: 10.1037/0033-2909.88.3.588
97. Bentler PM. Comparative fit indexes in structural models. *Psychol Bull.* (1990) 107:238–46. doi: 10.1037/0033-2909.107.2.238
98. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Model Multidiscip J.* (1999) 6:1–55. doi: 10.1080/10705519909540118
99. Jöreskog KG, Sörbom D. *LISREL 8.54. Structural Equation Modeling With the Simplis Command Language* (2004) Available online at: <http://www.unc.edu/~rcm/psy236/holzcf.lisrel.pdf>
100. Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS, editors. *Testing Structural Equation Models*. Newbury Park, CA: Sage Publications (1993). p. 136–62.
101. Yu C. *Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models With Binary and Continuous Outcomes*. Los Angeles, CA: University of California, Los Angeles. (2002).
102. Gerbing DW, Hamilton JG. Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Struct Equ Model Multidiscip J.* (1996) 3:62–72. doi: 10.1080/10705519609540030
103. Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Qual Life Res.* (2007) 16:19–31. doi: 10.1007/s11136-007-9183-7
104. Gibbons RD, Hedeker DR. Full-information item bi-factor analysis. *Psychometrika* (1992) 57:423–36. doi: 10.1007/BF02295430
105. Reise SP, Moore TM, Haviland MG. Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *J Pers Assess.* (2010) 92:544–59. doi: 10.1080/00223891.2010.496477
106. Brunner M, Nagy G, Wilhelm O. A Tutorial on hierarchically structured constructs. *J Pers.* (2012) 80:796–846. doi: 10.1111/j.1467-6494.2011.00749.x
107. Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research - Robert J. Vandenberg, Charles E. Lance, 2000. *Organ Res Methods* (2000) 3:4–70. doi: 10.1177/109442810031002

108. Sideridis GD, Tsaousis I, Al-harbi KA. Multi-population invariance with dichotomous measures: combining multi-group and MIMIC methodologies in evaluating the general aptitude test in the arabic language - Georgios D. Sideridis, Ioannis Tsaousis, Khaleel A. Al-harbi, 2015. *J Psychoeduc Assess*. 33:568–84. doi: 10.1177/0734282914567871
109. Joreskog K. A general method for estimating a linear equation system. In: Goldberger AS, Duncan OD, editors. *Structural Equation Models in the Social Sciences*. New York, NY: Seminar Press (1973). pp. 85–112.
110. Kim ES, Cao C, Wang Y, Nguyen DT. Measurement invariance testing with many groups: a comparison of five approaches. *Struct Equ Model Multidiscip J*. (2017) 24:524–44. doi: 10.1080/10705511.2017.1304822
111. Muthén B., Asparouhov T. *BSEM Measurement Invariance Analysis*. (2017). Available online at: <https://www.statmodel.com/examples/webnotes/webnote17.pdf>
112. Asparouhov T, Muthén B. Multiple-group factor analysis alignment. *Struct Equ Model*. 21:495–508. doi: 10.1080/10705511.2014.919210
113. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol Bull*. (1993) 114:552–66. doi: 10.1037/0033-2909.114.3.552
114. Pushpanathan ME, Loftus AM, Gasson N, Thomas MG, Timms CE, Olaithe M, et al. Beyond factor analysis: multidimensionality and the Parkinson's disease sleep scale-revised. *PLoS ONE* (2018) 13:e0192394. doi: 10.1371/journal.pone.0192394
115. Armor DJ. Theta reliability and factor scaling. *Sociol Methodol*. (1973) 5:17–50. doi: 10.2307/270831
116. Porta M. *A Dictionary of Epidemiology*. New York, NY: Oxford University Press (2008).
117. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* (1951) 16:297–334. doi: 10.1007/BF02310555
118. Zumbo B, Gadermann A, Zeisser C. Ordinal versions of coefficients alpha and theta for likert rating scales. *J Mod Appl Stat Methods* (2007) 6:21–9. doi: 10.22237/jmasm/1177992180
119. Gadermann AM, Guhn M, Zumbo B. Estimating ordinal reliability for Likert type and ordinal item response data: a conceptual, empirical, and practical guide. *Pract Assess Res Eval*. (2012) 17:1–13. Available online at: <http://www.pareonline.net/getvn.asp?v=17&n=3>
120. McDonald RP. *Test Theory: A Unified Treatment*. New Jersey, NJ: Lawrence Erlbaum Associates, Inc (1999).
121. Revelle W. Hierarchical cluster analysis and the internal structure of tests. *Multivar Behav Res*. (1979) 14:57–74. doi: 10.1207/s15327906mbr1401_4
122. Revelle W, Zinbarg RE. Coefficients alpha, beta, omega, and the glb: comments on Sijtsma. *Psychometrika* (2009) 74:145. doi: 10.1007/s11336-008-9102-z
123. Bernstein I, Nunnally JC. *Psychometric Theory*. New York, NY: McGraw-Hill (1994).
124. Weir JP. JP: Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Con Res*. (2005) 19:231–40. doi: 10.1519/15184.1
125. Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Stat Med*. (2002) 21:3431–46. doi: 10.1002/sim.1253
126. Churchill GA. A paradigm for developing better measures of marketing constructs. *J Mark Res*. (1979) 16:64–73. doi: 10.2307/3150876
127. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med*. (1990) 20:337–40. doi: 10.1016/0010-4825(90)90013-F
128. Hebert JR, Miller DR. The inappropriateness of conventional use of the correlation coefficient in assessing validity and reliability of dietary assessment methods. *Eur J Epidemiol*. (1991) 7:339–43. doi: 10.1007/BF00144997
129. McPhail SM. *Alternative Validation Strategies: Developing New and Leveraging Existing Validity Evidence*. San Francisco, CA: John Wiley & Sons, Inc (2007).
130. Dray S, Dunsch F, Holmlund M. *Electronic Versus Paper-Based Data Collection: Reviewing the Debate*. The World Bank Development Impact (2016). Available online at: <https://blogs.worldbank.org/impactevaluations/electronic-versus-paper-based-data-collection-reviewing-debate> (Accessed November 10, 2017).
131. Ellen JM, Gurvey JE, Pasch L, Tschann J, Nanda JP, Catania J. A randomized comparison of A-CASI and phone interviews to assess STD/HIV-related risk behaviors in teens. *J Adolesc Health* (2002) 31:26–30. doi: 10.1016/S1054-139X(01)00404-9
132. Chesney MA, Neilands TB, Chambers DB, Taylor JM, Folkman S. A validity and reliability study of the coping self-efficacy scale. *Br J Health Psychol*. (2006) 11(Pt 3):421–37. doi: 10.1348/135910705X53155
133. Thurstone L. *Multiple-Factor Analysis*. Chicago, IL: University of Chicago Press (1947).
134. Fan X. Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educ Psychol Meas*. (1998) 58:357–81. doi: 10.1177/0013164498058003001
135. Glockner-Rist A, Hoijtink H. The best of both worlds: factor analysis of dichotomous data using item response theory and structural equation modeling. *Struct Equ Model Multidiscip J*. (2003) 10:544–65. doi: 10.1207/S15328007SEM1004_4
136. Keeves JP, Alagumalai S, editors. *Applied Rasch Measurement: A Book of Exemplars: Papers in Honour of John P. Keeves*. Dordrecht ; Norwell, MA: Springer (2005).
137. Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures. *Clin Ther*. (2014) 36:648–62. doi: 10.1016/j.clinthera.2014.04.006
138. Harvey RJ, Hammer AL. Item response theory. *Couns Psychol*. (1999) 27:353–83. doi: 10.1177/0011000099273004
139. Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Qual. Life Res*. (2009) 18:447–60. doi: 10.1007/s11136-009-9464-4
140. Greca AML, Stone WL. Social anxiety scale for children-revised: factor structure and concurrent validity. *J Clin Child Psychol*. (1993) 22:17–27. doi: 10.1207/s15374424jccp2201_2
141. Frongillo EA, Nanama S, Wolfe WS. *Technical Guide to Developing a Direct, Experience-Based Measurement Tool for Household Food Insecurity*. Washington, DC: Food and Nutrition Technical Assistance Project (2004).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Boateng, Neilands, Frongillo, Melgar-Quinonez and Young. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.