Measure Development

Research Design & Analysis Core

Version: 2.0 • 02 May 2022

Duke GLOBAL HEALTH INSTITUTE | Research Design & Analysis Core

Contents

Introduction 01

Steps for Development 04

Reliability 06

Validity 09

Power and Sample Size 16

References 17

Introduction

In many studies, researchers desire to measure an intangible concept or construct. Unlike body weight, for example, which can be measured directly, these concepts are not observed and, thus, cannot be measured directly. Examples of these unobserved variables (known more commonly in the statistical literature as “latent variables”) abound, especially in psychology and the social sciences. Mental health issues and personality disorders, for example, are not generally directly measurable, but must be inferred from questions asked of the participant about behaviors or feelings. Usually, several questions must be asked to get at the underlying concept. For example, in political science, “conservatism” and “liberalism” cannot be directly measured, but may be inferred by answers to several questions.

To help illustrate key concepts in measurement and latent variables, we will use depression as an example. While a single question, such as “Do you commonly feel blue?” may purport to measure depression, it is clearly not a good measure of depression (we will discuss what constitutes “good” in the sections below). First, while an answer of “yes” may identify most people who are genuinely depressed, it will also likely include people who have a melancholy personality or are going through a tough time because of a recent loss, but who are not clinically depressed. Second, as indicated by the clinical definition in the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) (American Psychiatric Association. & American Psychiatric Association. DSM-5 Task Force., 2013), depression is a multifaceted illness which includes more than “feeling sad”. With only one question, we may be missing important facets of depression, and thus fail to identify individuals with depression. In fact, a gold standard for measuring depression is the structured clinical interview for DSM disorders (SCID), in which 10 different items related to depression are evaluated by a trained psychiatrist, and if at least 5 out of 10 of those items are endorsed, the person is considered depressed (Gallis et al., 2018; Spitzer et al., 1992).

The measurement of an unobserved (latent) variable can be illustrated using the patient health questionnaire (PHQ-9), which is commonly used to measure depression (Kroenke et al., 2001). **Figure 1** displays the concept of a latent variable in graph format. The PHQ-9 has nine questions (items) related to depression, with four answers for each question ranging from "Not at all" (coded as 0) to "Nearly every day" (coded as 3). These nine items are observed variables—that is, we directly observe the respondent's answer to them. Once we have answers to the nine items, one way to find the total score is by summing them up, which gives us the "PHQ-9 score". As an alternative to a simple sum of the items, we could create a PHQ-9 score by creating a weighted sum of the items, estimated using a technique called factor analysis. (Grice, 2001). Since the PHQ-9 is intended to measure depression, we label the latent variable "depression". Depression is not directly observed but instead is measured as a (weighted or unweighted) combination of nine observed survey responses.

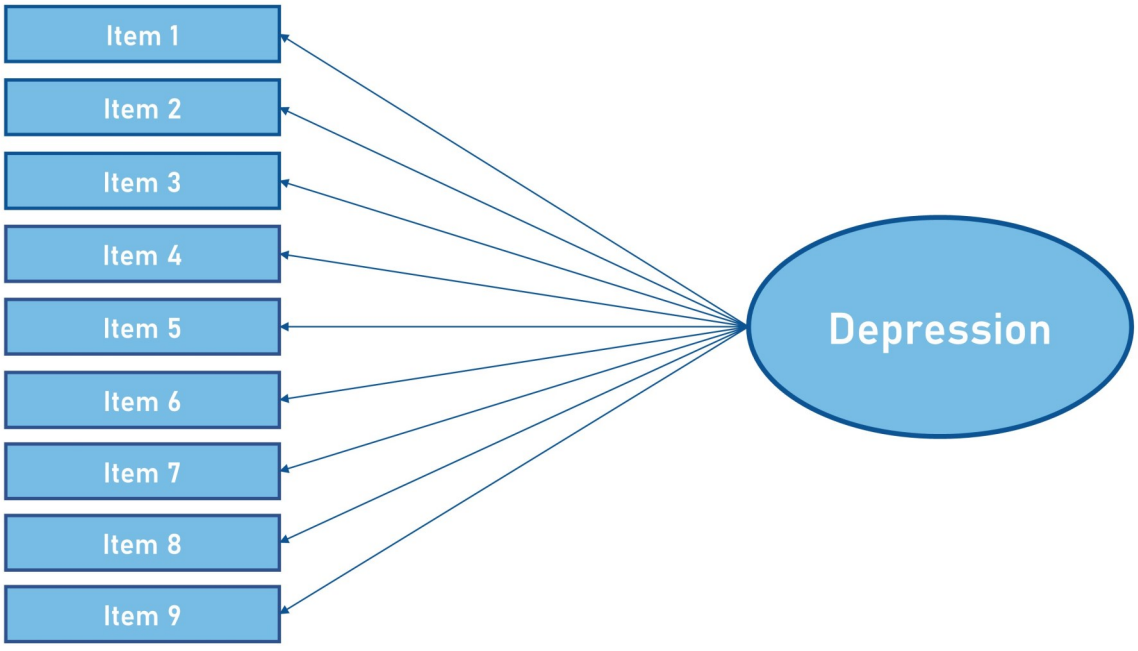

Figure 1. Illustration of the path diagram of Depression as measured by the PHQ-9 instrument.

Often, the construct a researcher wishes to measure does not have a scale available to measure it, or the scale exists but was developed in a different language or for a different population. For example, a researcher may wish to measure *stigma* (latent variable) related to an illness for which a stigma scale does not already exist. Or, it may be that the researchers want to measure a novel personality construct. In the case of developing a novel latent construct, or the adaptation of an existing construct to a new cultural setting, researchers will require statistical methods to validly measure the latent variable. Additionally, a scale may need to be adapted if used in a different cultural setting, including replacing words or items with more culturally appropriate items or translating the instrument into a new language.

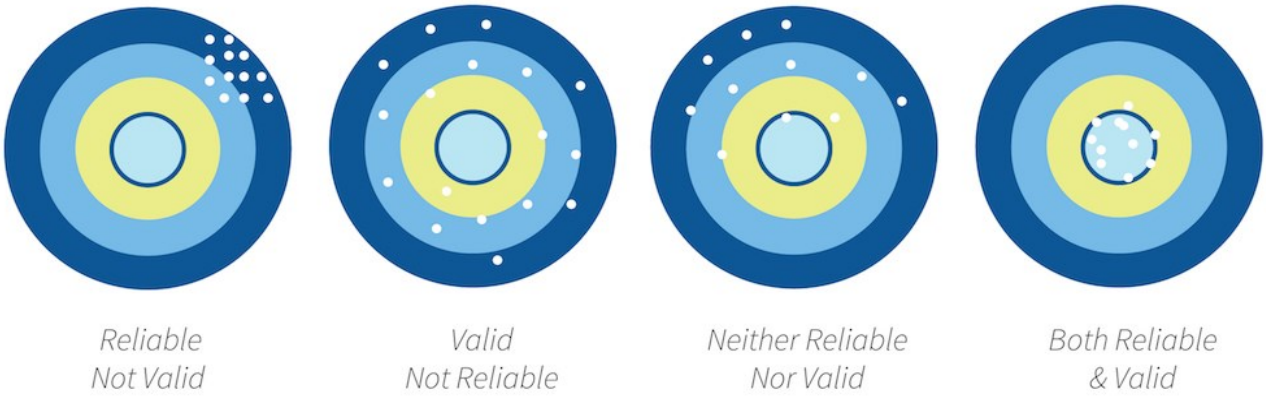

Figure 2. Visual representation of validity and reliability, copied from www.healthysimulation.com/16389/understanding-research-for-clinical-simulation-part-2-validity-and-reliability/

Measurement theory has been widely developed with several different approaches. This guide will cover two main concerns when creating a new measure of a construct: to ensure validity and to ensure reliability. Validity is concerned with the question: "Is this construct measuring what we intended?" A measure to identify depression is no good for its purpose if it actually measures schizophrenia. Reliability asks the questions: "Is the measure consistent? Does it produce the same readings in the same circumstances?" (Coolican, 2014, p. 40) These concepts are illustrated conceptually in **Figure 2**. We will discuss these concepts and how to measure them in the sections below.

Steps for Development

1. Define what should be measured

An important first step in scale development is to define what we would like to measure. This involves conceptualizing the concept (latent variable), soliciting input from experts in the field, and reading literature in order to theoretically define what we are measuring.

2. Creating or Adapting Scales

Once we have decided on which construct we'd like to measure, we may either

1) Use an existing scale if it has already been validated for our population of interest
2) Adapt an existing scale measuring that construct
3) Create our own brand new scale measuring the construct
4) Some combination of the three.

This part of scale development is primarily qualitative, and is often time-intensive. First, in-depth interviews are conducted with a small number of people from the population in which the scale will be evaluated, as well as key stakeholders, such as healthcare providers and government ministries. These interviews are then summarized, and often semi-structured interview questions are updated after each interview to inform subsequent interviews. After this, a technique such as thematic analysis may then be used to identify and evaluate themes within the interviews, further refining which questions should be used to measure the construct (Chapman et al., 2015).

This information is then used to develop an item pool (that is, a set of potential items for the new scale). If adapting an existing scale, items may come entirely or partly from that scale. Otherwise, item content is generated from the qualitative interviews. Based on qualitative data, items from an existing scale may be adapted. If items from the existing scale to be

adapted are not in the language of interest for the current study, they will be translated and back-translated to ensure a culturally appropriate translation (Brislin, 1970). For example, if a depression screener to be translated from English to Urdu contains the phrase “feeling blue”, translation and back-translation would help ensure that the translated scale uses a culturally and linguistically similar phrase to help get at the same concept.

Once this step is complete, we will have a pool of items creating a scale that is purported to measure what we wish to measure. Next, we will recruit a sufficient sample to perform a validity and reliability study of the measure (see the section *Power and Sample Size* for guidance on selecting a “sufficient” sample). It is this phase we turn to in the following sections.

How do we measure reliability and validity? Anthoine et al. (2014) have a helpful table displaying different types of validity and reliability that may be measured. We have adapted this table as a useful reference (see **Table 1** and **Table 2**), but for a more exhaustive list, see Anthoine et al. (2014). We will discuss some of these measures in the following sections.

Frost et al. (2007) state that “reliability is necessary but not sufficient in determining validity.” But “measures can be highly reliable but not measure what they are purported to measure” (i.e., not valid; see Figure 2, panel 1). Thus, it is necessary to determine both validity and reliability.

Reliability

Reliability has been classically divided into two concepts: internal and external. These are defined and explained in the following table and subsections.

Table 1: Concepts in Measure Design: Reliability

Concept	APA Standard	Brief Definition	How to Measure	Example using Depression
Internal Reliability	Internal consistency	"Provides information about the associations among different items in the scale" (Frost et al., 2007).	Computing a statistic, such as Cronbach's alpha or coefficient omega, or using the **split-half method**.	Compute alpha on your new depression measure. A rule of thumb is that good internal consistency is an alpha of between 0.7 and 0.95.
External Reliability	Alternate form and test-retest	"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" (Anthoine et al., 2014), or when using different forms of the same questionnaire.	**Test-retest method** where participants are tested on the same instrument a short period of time apart. **Agreement or concordance coefficients** to show the similarity of different forms of administration.	Administer your new depression measure, and then re-administer it 2-14 days later. Use ICC or Kappa to measure concordance.

Internal Reliability

Internal reliability "provides information about the associations among different items in the scale" (Frost et al., 2007). The most common statistic used to measure this is Cronbach's alpha. Cronbach's alpha is a measure of internal consistency assessed by the average inter-correlation among the items (and hence reliability). Cronbach's alpha of scale items increases as the number of items in the scale increases and as the intraclass correlation (ICC) among items increases. Although alpha is widely used, coefficient omega is sometimes recommended instead as a more general measure of reliability, which remains unbiased in situations where alpha is biased (Padilla & Divers, 2016). Another method for testing internal reliability is the **split-half method.** In this method, we divide the items randomly into two halves. Then, if the test is reliable "people's scores on each half should be similar", and this is generally determined by a Spearman-Brown correlation of 0.75 or higher (Coolican, 2014, pp. 216-217).

Cronbach's alpha is simple to compute using any statistical software package. Although several different criteria have been proposed for what constitutes "good" internal consistency, an alpha between about 0.70 and 0.95 is generally accepted to be "good" (Terwee et al., 2007). Note that if alpha is above 0.95, this may indicate too much redundancy among questions. Cronbach's alpha can also be artificially inflated by adding many questions to the scale.

External Reliability

External reliability is generally measured using the test-retest method, which is "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" (Anthoine et al., 2014). Often, test-retest reliability will be determined by re-administering the survey to the whole sample or a subset of the sample a short period of time after first administration. This time period should be chosen to be short enough that the underlying construct has not changed, but long enough that the respondents will not fully remember their previous answers to the questions. Typically, an interval of 2 to 14 days is recommended (Streiner et al., 2015).

When using test-retest reliability to examine external reliability (reproducibility), we may compare the two measurements on participants using ICC (for continuous measures) or a weighted Kappa (for ordinal measures). Citing Nunnally and Bernstein (1994), Terwee et al. (2007) consider an ICC or weighted Kappa of at least 0.70 to indicate good external reliability, as long as the sample size is at least 50.

External reliability might also be related to the use of alternate forms of the same instrument. For example, questionnaires are commonly reported in long and short versions, with differing number of questions. Reliability should be tested independently for each version. A similar notion should be applied to different forms of administration of a questionnaire. Theophanous et al. (2019) recently showed that the evaluation of snakebite envenomation patients' functional assessment, using the Patient Specific Functional Scale, was reliable when administered in-person and through the telephone. This same concept applies to smartphone vs. paper-based questionnaires, or different language versions of the same instrument.

Validity

Validity is “the extent to which an instrument measures what it was intended to measure and not something else” (Frost et al., 2007). Both panels B and C of figure 2 display validity in visual form. Frost et al. (2007) define “three main subtypes of validity: content, criterion, and construct.” They state that “a strong correlation should be demonstrated with measures addressing similar constructs and a weak correlation with measures addressing disparate constructs.” Further types of validity are described in the APA standards (American Educational Research et al., 2014), and discussed below.

Table 2: Concepts in Measure Design: Validity

Concept	APA Standard	Brief Definition	How to Measure	Example using Depression
Content Validity	Test-content	“The extent to which the instrument “measures the appropriate content and represents the variety of attributes that make up the measured construct” (Frost et al., 2007).	Qualitative measures such as focus groups or cognitive interviews.	When creating your new depression scale, you conduct focus groups or individual interviews with experts in the field, mental health professionals, and patients to ask about their comprehension and perceptions of cultural relevance of the questions.
Construct Validity	Internal structure	The “extent to which the measure ‘behaves’ in a way consistent with theoretical hypotheses” (Frost et al., 2007).	Methods such as principal compo-nent analysis, factor analysis, and item re-sponse theory.	You analyze participants' responses data using factor analysis to identify the underlying dimensionality structure, or latent model.

Table 2 continued: Concepts in Measure Design: Validity

Concept	APA Standard	Brief Definition	How to Measure	Example using Depression
Criterion-Related Validity/ Concurrent Validity	Relationship with other variables	The "extent to which the measure agrees with an external standard measure" (Frost et al., 2007).	Comparing the new scale to a reference standard which is purported to measure a similar domain. If the reference standard is considered a "gold standard", this type of validity is referred to as **diagnostic validity.**	Compare scale cutoffs of your new depression measure with the structured clinical interview for depression (SCID; the gold standard of measuring depression). Use sensitivity and specificity to determine cutoff to indicate depression on your new scale. See Gallis et al. (2018) for an example.
Face Validity	Test-content	Whether a scale *appears* to measure what it purports to measure.	Experts in the field provide feedback on the items.	Solicit feedback from psychologists about your new depression measure.
N/A	Response process	"The fit between the construct and the detailed nature of the performance or response actually engaged in by test takers" (American Educational Research et al., 2014)	Open ended question about the rationale of the question or think-aloud responses during a focus group session.	Measure respondents' understanding of the questions on your new depression measure using open-ended questions.
N/A	Consequence of testing	The "soundness of proposed interpretations [of test scores] for their intended uses" (American Educational Research et al., 2014).	Surveys about the usage of the test response in practice. Understand how the score will impact practice or the social context.	

Test Content

Validity regarding test-content is the extent to which the instrument "measures the appropriate content and represents the variety of attributes that make up the measured construct" (Frost et al., 2007). This form of validity can be informed by qualitative methods such as in-depth interviews, focus groups, and cognitive interviews. The idea of cognitive interviews was developed by researchers in psychology who applied their knowledge of how memory works to help people recall events (Bull, 1996). Empirical studies confirmed that cognitive interviews "generated approximately 30% to 35% more correct information [compared to standard interviews,] without increasing the number of incorrect or confabulated details" (Bull, 1996).

In addition, experts in the field can examine the items to determine **content validity (also referred to as face validity)** —that is, whether the scale *appears* to be measuring what it intends to measure. Currently, instead of using expert opinions, researchers generally focus on judges' evaluation to provide input about the content of a given instrument. Judges could be experts in the field but could also be experienced practitioners or community members. Quantitative measures can be used to verify test content, with indicators such as the content validity coefficient or the kappa coefficient for the agreement between judges.

Response Process

Validity regarding the response process is the extent to which the targeted theoretical construct fits the questionnaire response process, performance, or the way respondents engage the test. For instance, a cognitive status assessment assumes a certain level of education by the response takers. As such, it has been widely reported in the literature that the education level influences the performance of test takers in cognitive assessments such as the MMSE or the MoCA questionnaires (Matallana et al., 2011). As education is a relevant

factor in the response of cognitive tasks, motor function might be relevant to questionnaires that involve drawing or writing. Therefore, in order to provide evidence of validity regarding response process, investigators should be concerned about the "theoretical and empirical analysis of the response process of test takers" (American Educational Research et al., 2014).

Individual evaluations of the response process could be used as metrics for this method of validity assessment. Focus groups to identify the reasoning about the response process, or indicators such as eye tracking or response times could also provide evidence about this type of validity. Investigations by judges or observers might also be relevant to ascertain the if the test is appropriately scored or applied.

Consequence of Testing

Validity regarding consequence of testing is defined as the extent to which the interpretations originated by the questionnaire response (or the scores obtained in the process) are sound in relation to the application they were intended to (American Educational Research et al., 2014). Messick (1995) goes further and include the assessment of the implications of the questionnaire scores/outputs and its social impacts. A classic example is the use of personality assessment to dictate job positions or students' entry to a university. As much as it is important to offer opportunities in which people can strive and reach their best, a personality assessment can also lead to the development of prejudice and discrimination.

Relationship with Other Variables

Validity regarding the relationship with other variables concerns the extent to which the measurement output related to variables external to the test itself (American Educational Research et al., 2014). Classically, there are three main formats of validity regarding the relationship with other variables: evidence regarding how a measure associated with (a)

another measure/variable of the same/similar construct (known as convergent validity), (b) a measure/variable of different or distinct constructs (known as discriminant validity), (c) an external to the construct measure (known as criterion validity).

Convergent and discriminant validity refer to the ability of the questionnaire to associate with other constructs. A study evaluating the evidence of validity of the PHQ-9 (a depression measure) could associate the measure with the results of HDRS (another depression scale). Another approach would be to see the ability of the PHQ-9 to differentiate the depression levels of patients with a known diagnosis of depression and people without a depression diagnosis (approach also referred to as known group comparison). The goal is to evaluate if the measure will behave as expected theoretically (American Educational Research et al., 2014).

Criterion-related validity "refers to the extent to which the measure agrees with an external standard measure" (Frost et al., 2007). This type of validity is commonly also referred to as **concurrent validity** (Coolican, 2014, p. 222), and it assumes there is a reference standard to compare the new scale against (Terwee et al., 2007). If there is such a "gold standard" reference measure, then measures such as area under the curve and sensitivity and specificity would be appropriate to report to assess **diagnostic validity** (Kohrt et al., 2011).

An example of criterion-related validity can be found in Gallis et al. (2018), in which the diagnostic validity of the PHQ-9 was compared to the gold standard of the structural clinical interview for depression (SCID). The SCID is a dichotomous measure indicating if the woman is depressed or not (Spitzer et al., 1992). The authors used Cronbach's alpha to determine internal reliability, and computed sensitivity and specificity at various cut-offs of the distribution to determine diagnostic validity. They determined that in the study population, the standard cutoff of PHQ-9 ≥ 10 worked well in screening for depression, finding that about 3 out of 4 women assessed positive for depression using the PHQ-9 cutoff of 10 also assessed positive on the SCID.

Internal structure

Construct validity is "the extent to which the measure 'behaves' in a way consistent with theoretical hypotheses" (Frost et al., 2007). Thus, for example, we would want the scale to be positively correlated with similar measures and negatively correlated (or uncorrelated) with dissimilar measures. "Construct validity is typically examined using bivariate correlations, factor analysis, and multivariate regression methods" (Frost et al., 2007). As Coolican (2014) points out, construct validity is "not a simple one-off procedure that will give us a numerical value for the validity of a scale. It is the development of evidence for a hypothetical psychological construct through the rigors of hypothesis testing and scientific method" (Coolican, 2014, p. 223).

Construct validity may be determined by examining how well the new scale correlates with an existing scale measuring a similar construct. Terwee et al. (2007) recommend that this is tested using predefined hypotheses, with a common test being whether the new scale has at least a 75% (Pearson) correlation with a similar existing scale. There are several approaches and a wide literature on the statistical methods used to evaluate internal structure of a measurement tool. The most common is factor analysis and its variations, but reference should also be made to item response theory and network analysis.

Exploratory and Confirmatory Factor Analysis

Factor analysis, part of construct validity, "is a useful analytic tool that can tell us…about important properties of a scale. It can help us determine *empirically* how many constructs [latent variables or *factors*] underlie a set of items" (DeVellis, 2012).

Factor analysis can either be exploratory or confirmatory. Exploratory factor analysis (EFA) is "a data-driven approach such that no specifications are made in regard to the number of common factors (initially) or the pattern of relationships between the common factors and the indicators (i.e., the factor loadings)" (Hoyle, 2012, p. 361). Thus, EFA is used to "determine the

appropriate number of common factors, and to ascertain which measured variables are reasonable indicators of the various latent dimensions" (Hoyle, 2012, pp. 361-362). In confirmatory factor analysis (CFA), on the other hand,

> the researcher specifies the number of factors … as well as other parameters, such as those bearing on the independence or covariance of the factors and indicator unique variances. The prespecified factor solution is evaluated in terms of how well it reproduces the sample covariance matrix of the measured variables. Unlike EFA, CFA requires a strong empirical or conceptual foundation to guide the specification and evaluation of the factor model. Accordingly, **EFA is often used early in the process of scale development whereas CFA is used in the later phases, when the underlying structure has been established on prior empirical and theoretical grounds** (Hoyle, 2012, p. 362).

Generally, factors should have at least three variables and item loadings of at least 0.32 (Yong & Pearce, 2013), although there is disagreement over this, with others stating that the item loadings should be at least 0.5 (Hair et al., 2018). We have not found a rule of thumb on the number of levels an item should have, but since items are supposed to be describing an underlying continuous variable, items with at least 5 levels (e.g., a 5-point Likert scale from 1= "strongly agree" to 5 = "strongly disagree") are recommended. In fact, there is a tradeoff between the number of item levels and number of items in a scale. Generally speaking, the fewer item levels (response options), the more items will be needed in the scale to maintain an acceptable level of reliability (Simms et al., 2019). The number of levels should also be defined theoretically according to the expected construct or behavior assessed. Methods such as taxonometric analysis help identify the metric system for a specific construct (Gordon et al., 2007).

For more details on Factor Analysis, see, for example, DeVellis (2017).

Power and Sample Size

In studies of reliability and validity of an instrument, there are no hard and fast statistical rules governing power and sample size. In their highly cited paper on the measurement properties of health status questionnaires, Terwee et al. (2007) consider a validity and reliability study to be of high quality if, among other things, the sample size is at least the maximum of 100 people or 7 people*(# of items)—that is, a minimum sample size of 100—assuming that the study is testing internal consistency using Cronbach's alpha. In another study, Frost et al. (2007) argue that 200 people is the minimum suggested for starting with the most basic psychometric analyses. Frost et al. (2007) also argue that "**replication** of psychometric estimates is needed either by a sufficiently large and representative sample that can be split into two subsamples for cross-validation or two samples of sufficient sample size." Others they cite recommend at least 300 people (and at least 5 people per variable) for factor analyses. Tabachnick and Fidell (2013, p. 618) state that "at least 300 cases are needed with low communalities [how much items correlate with one another], a small number of factors, and just three or four indicators for each factor", although if there are more than four indicators and the factors are well-determined (loadings > 0.8), a sample size of 100-200 is probably acceptable.

CONCLUSION

This has been a brief introduction to scale development. We have not discussed the entire breadth and depth of the field; there are other theories and approaches to scale development. The interested reader is referred to excellent references such as DeVellis (2017), Frost et al. (2007), or Terwee et al. (2007) for greater detail.

References

American Educational Research, A., American Psychological, A., National Council on Measurement in, E., Joint Committee on Standards for, E., & Psychological, T. (2014). *Standards for educational and psychological testing*. Washington, DC.

American Psychiatric Association., & American Psychiatric Association. DSM-5 Task Force. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5 (*5th ed.). Washington, D.C.: American Psychiatric Association.

Anthoine, E., Moret, L., Regnault, A., Sébille, V., & Hardouin, J.-B. (2014). Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. *Health Qual Life Outcomes,* 12, 2. doi:10.1186/s12955-014-0176-2

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology,* 1(3), 185-216.

Bull, P. R. (1996). The cognitive interview: What is it and does it work? *Child Care in Practice*, 2 (3), 59-66. doi:10.1080/13575279608415290

Chapman, A., Hadfield, M., & Chapman, C. (2015). Qualitative research in healthcare: an introduction to grounded theory using thematic analysis. *Journal of the Royal College of Physicians of Edinburgh,* 45(3), 201-205.

Coolican, H. (2014). *Research methods and statistics in psychology* (Sixth edition. ed.). London; New York: Psychology Press, Taylor & Francis Group.

DeVellis, R. F. (2012). *Scale development : theory and applications* (3rd ed.). Thousand Oaks, Calif.: SAGE.

DeVellis, R. F. (2017). *Scale development : theory and applications* (Fourth edition. ed.). Los Angeles: SAGE.

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D., & Group, M. F. P.-R. O. C. M. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health*, 10, S94-S105.

Gallis, J. A., Maselko, J., O'Donnell, K., Song, K., Saqib, K., Turner, E. L., & Sikander, S. (2018). Criterion-related validity and reliability of the Urdu version of the patient health questionnaire in a sample of community-based pregnant women in Pakistan. *PeerJ,* 6, e5185. doi:https://10.7717/peerj.5185

Gordon, K., Holm-Denoma, J., Smith, A., Fink, E., & Joiner, T., Jr. (2007). Taxometric analysis: introduction and overview. *International Journal of Eating Disorders*, *40 Suppl,* S35-39. doi:10.1002/eat.20407

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods,* 6(4), 430-450. doi:10.1037/1082-989X.6.4.430

Hair, J. F., Babin, B. J., Anderson, R. E., & Black, W. C. (2018). *Multivariate data analysis* (Eighth edition. ed.). Andover, Hampshire: Cengage.

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. New York: Guilford Press.

Kohrt, B. A., Jordans, M. J. D., Tol, W. A., Luitel, N. P., Maharjan, S. M., & Upadhaya, N. (2011). Validation of cross-cultural child mental health and psychosocial research instruments: adapting the Depression Self-Rating Scale and Child PTSD Symptom Scale in Nepal. *BMC Psychiatry*, 11, 127-127. doi:10.1186/1471-244X-11-127

Kroenke, K., Spitzer, R., & Williams, J. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606-613.

Matallana, D., de Santacruz, C., Cano, C., Reyes, P., Samper-Ternent, R., Markides, K. S., ... Reyes-Ortiz, C. A. (2011). The relationship between education level and mini-mental state examination domains among older Mexican Americans. *Journal of Geriatric Psychiatry and Neurology,* 24(1), 9-18. doi:10.1177/0891988710373597

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational measurement: Issues and practice,* 14(4), 5-8.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Padilla, M. A., & Divers, J. (2016). A comparison of composite reliability estimators: coefficient omega confidence intervals in the current literature. *Educational and Psychological Measurement*, 76(3), 436-453.

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, 31(4), 557.

Spitzer, R. L., Williams, J. B., Gibbon, M., & First, M. B. (1992). The structured clinical interview for DSM-III-R (SCID): I: history, rationale, and description. *Archives of General Psychiatry*, 49 (8), 624-629.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: a practical guide to their development and use*: Oxford University Press, USA.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson Education.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., . . . de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, 60(1), 34-42.

Theophanous, R. G., Vissoci, J. R. N., Wen, F. H., Griffin, S. M., Anderson, V. E., Mullins, M. E., . . . Gerardo, C. J. (2019). Validity and reliability of telephone administration of the patient-specific functional scale for the assessment of recovery from snakebite envenomation. *PLoS Neglected Tropical Diseases,* 13(12), e0007935-e0007935. doi:10.1371/journal.pntd.0007935

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology,* 9 (2), 79-94.

Prepared by: Joao Vissoci, John Gallis, Rachel D'Anna
Reviewed by: Duke Global Health Institute | Research Design & Analysis Core
Cover photo by: Photo by Dan-Cristian Păduret on Unsplash