Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 26 Number 1, January 2021 ISSN 1531-7714

A Practical Guide to Instrument Development and Score Validation in the Social Sciences: The MEASURE Approach

Michael T. Kalkbrenner, *New Mexico State University*

The research and practice of social scientists who work in a myriad of different specialty areas involve developing and validating scores on instruments as well as evaluating the psychometric properties of existing instrumentation for use with research participants. In this article, the author introduces The MEASURE Approach to instrument development, an acronym of seven empirically supported steps for instrument development, and initial score validation that he developed based on the recommendations of leading psychometric researchers and based on his own extensive background in instrument development. Implications for how The MEASURE Approach has utility for enhancing the assessment literacy of social scientists who work in a variety of different specialty areas are discussed.

Introduction

Assessment literacy is a pertinent issue in social sciences research, as researchers tend to assess latent variables (e.g., personality, morale, and other attitudinal variables) that are abstract in nature, which are generally appraised by inventories (Gregory, 2016). To this end, social science researchers and practitioners are responsible for understanding the basic foundations, operations, and applications of testing, including instrument development (Standards for Educational and Psychological Testing, 2014). Instrumentation with strong psychometric support, however, tends to be underutilized by social scientists when conducting program evaluation and other types of research (Tate et al., 2014). The extant literature includes a series of peer-reviewed journal articles (e.g., Benson, 1998; Kane, 1992; Mvududu & Sink, 2013), as well as textbooks and book chapters (e.g., Bandalos & Finney, 2019; DeVellis, 2016; Dimitrov, 2012; Fowler, 2014; Gregory, 2016; Kane & Bridgeman, 2017), which collectively outline the instrument development process as well as guidelines for testing the validity of inferences from the scores. However purchasing these resources is infeasible for many graduate students, who are already financially burdened by the cost of required textbooks. Similarly, applied social scientists (e.g., counselors and teachers) who are not affiliated with a university that provides access to electronic data bases might also have limited access to these resources.

The literature is lacking a single refereed journal article (a one-stop-shop) that includes a practical outline of the instrument development and validation of scores process based on a number of synthesized recommendations of prominent expert, contemporary psychometric researchers. Such an article has potential to provide social scientists with a single and accessible resource for developing their own measures as well as for evaluating the rigor of existing measures for use with research participants. The primary aim of the present author was to introduce The MEASURE Approach to instrument development. MEASURE (Figure 1) is an acronym comprised of the first letter of the following seven empirically supported steps for developing and validating scores on measures: (a) **M**ake the purpose and rationale clear, (b) **E**stablish empirical framework, (c) **A**rticulate theoretical blueprint, (d) **S**ynthesize content and scale development, (e) **U**se expert reviewers, (f) **R**ecruit participants, and (g) **E**valuate validity and reliability.

The MEASURE Approach was developed based on the guidelines of leading psychometricians, primarily Benson (1998), DeVellis (2016), Dimitrov (2012), and Mvududu and Sink (2013), as well as the author's extensive background and experience with instrument development and score validation. Finally, an exemplar description of an instrument development study conducted by Kalkbrenner and Gormley (2020) is presented to provide an example of applying each step in The MEASURE Approach.

Figure 1. The MEASURE Approach to Instrument Development

Make the purpose and rationale clear

Establish empirical framework

Articulate theoretical blueprint

Synthesize content and scale development

Use expert reviewers

Recruit participants

Evaluate validity and reliability

Step 1: Make the Purpose and Rationale Clear

Researchers should first define the purpose of conducting an instrument development study by telling the reader what they are seeking to measure and why measuring the proposed construct is important (DeVellis, 2016; Dimitrov, 2012). As part of this step, researchers should review the existing literature on the proposed construct of measurement to determine if they can use/adapt an existing measure or if an instrument development study is necessary (Mvududu & Sink, 2013). If a measure exists in the literature, researchers should carefully evaluate the rigor of the instrument development study by comparing the procedures that the test developers employed to established empirical standards (e.g., The MEASURE Approach). An instrument development study is necessary if the literature is lacking a measure to appraise the researcher's desired construct of measurement. An instrument development study might also be necessary if a researcher determines that an existing instrument is potentially psychometrically flawed (e.g., lacking reliability or validity evidence, step 7).

An instrument development study might also be necessary if a researcher determines that an existing instrument is inappropriate for use with their target population (e.g., cross-cultural fairness issues). In some instances, developing an original measure with a diverse population can be more appropriate than confirming scores on an established measure (step 7) that was developed and normed with a different population. Suppose for example, a researcher is seeking a screening tool for appraising mental health distress among Spanish speaking clients. There might be utility in creating a new screening tool (based on the culture) rather than trying to validate scores on an existing measure with Spanish speaking clients, as the nature and breadth of the construct of measurement (content validity, step 2) can vary substantially between different cultures. Thus, even if an existing measure of mental health distress is found to be statistically sound with Spanish speaking clients, it might fail to capture unique elements of mental health distress in the culture (see Kane, 2010, for an overview of fairness-related considerations in testing and assessment). After making the purpose clear, a researcher should provide a rationale to justify why creating a new instrument is necessary.

When providing a rationale for developing a new instrument, researchers should (a) present a summary of their review of the extant measurement literature, (b) cite any similar instruments that already exist, and (c) articulate the construct(s) that existing measures fail to capture in order to highlight a gap in the existing measurement literature (DeVellis, 2016; Dimitrov, 2012). Finally, test developers should discuss how their proposed instrument has potential to fill the aforementioned gap in the measurement literature and articulate how filling this gap has significant potential to advance future research and practice (see Fu and Zhang, 2019, as well as Kalkbrenner and Gormley, 2020, for examples of providing a rationale for instrument development based on these steps).

Step 2: Establish Empirical Framework

Benson (1998) suggested instrument development undergoes a Substantive Stage in which test developers situate the study within the context of a theoretical framework. Similarly, in the Establish Empirical Framework stage, researchers are tasked with identifying a theory(ies) and/or synthesized findings from the extant literature to set an empirical framework for the item development process. In this context, an empirical framework refers to at least one theory or scholarly source (e.g., peer-reviewed) that provides a series of principles or assumptions that underlie the proposed construct of measurement. For example, a test developer might refer to Maslow's Hierarchy of Needs (Maslow, 1943) as the empirical framework for developing a measure to appraise the extent to which one's various needs are satisfied. The goal in step 2 is to provide an overview of the theoretical underpinnings for the proposed construct of measurement, which is an important step for ensuring content validity or the extent to which test items adequately represent the scope of a construct of measurement (Lambie et al., 2017). Four primary methods for demonstrating content validity in social sciences research include (a) empirical framework, (b) theoretical blue print (step 3), (c) expert review (step 5), and (d) pilot testing (step 6).

In some instances, the literature might be lacking an established theory that a researcher can use to set an empirical framework for the item development process. In these instances, a researcher can build their own theoretical framework for item development by synthesizing the findings from a number of empirical sources (e.g., peer-reviewed journal articles) that collectively provide a rationale for the intended construct of measurement. At this stage of development, the empirical framework can be general in nature. The idea is to refer to at least one empirical source that will set the framework for developing items that capture the proposed construct of measurement.

Step 3: Articulate Theoretical Blueprint

Researchers can begin to refine and organize their empirical framework by creating a theoretical blueprint. A theoretical blueprint (Figure 2) is a tool for enhancing the content validity of a measure by offering researchers two primary advantages, including (a) creating the content and domain areas for the construct of measurement and (b) determining the approximate proportion of items that should be developed across each content and domain area (Menold et al., 2015; Summers & Summers, 1992). Content areas in a blueprint refer to the specific subject aspects for the construct of measurement. Domain areas in a blueprint refer to the various application-based dimensions of the construct of measurement. The content and domain areas on a blueprint should be derivatives of the extant literature and, in most cases, multiple plausible/logical content and domain areas can be generated for a construct of measurement; thus, there is usually not one "right" or "correct" content or domain area for any given measure. Researchers are tasked with providing a rationale from the extant literature to justify the utility of their content and/or domain areas.

Figure 2. Example Theoretical Blueprint: Mental and Physical Health (Kalkbrenner & Gormley, 2020)

		Domain Areas		
		Frequency	**Intensity**	**Duration**
Content Areas	**Diet**	7	6	6
	Exercise	7	6	6
	Stress Management	12	9	9
	Avoiding Toxins	4	4	4

Researchers should refer to the extant literature (step 1) and the empirical framework (step 2) to determine the breadth of their proposed construct of measurement. Researchers can also seek assistance from content experts (step 5) to help with item development and determining the breadth of their proposed construct of measurement. Seeking input from a panel of content experts who are representative of the field of study might be especially helpful in cases where there is a gap in literature on the proposed construct of measurement. To enhance content validity, test developers should adjust for the relative importance of the items across each content and domain area for the construct of measurement. In other words, more items should be developed for the intersecting content and domain areas that represent a greater scope of the construct of measurement. The purpose of numbering the intersecting cells on a blueprint (Figure 2) is to denote the approximate proportion of items that will be developed to represent each cell. Not every instrument, however, will be based on a theoretical framework that lends itself to the blueprint matrix that is depicted in Figure 2. As such, a test developer might include only content area(s) or only domain area(s) on their blueprint. Blueprint construction is a flexible procedure, which allows researchers to customize this tool to enhance content validity in the subsequent item development process.

Step 4: Synthesize Content and Scale Development

Synthesize Content

Before developing an initial pool of items, researchers should be clear about the parameters of their proposed construct of measurement and reflect on how their construct differs from other latent variables in order to avoid redundancy (DeVellis, 2016; Fowler, 2014). The empirical framework (step 2) and blueprint (step 3) can be instrumental tools to synthesize content for the purpose of refining the parameters of the construct of measurement during the item development process. Researchers should develop approximately three to four times as many items that will comprise the final version of the measure (DeVellis, 2016) as multiple (potentially problematic) items are usually deleted during the expert review (step 5) and during reliability/validity testing (step 7). Items should be brief, clear, and written at approximately a sixth-grade reading level (see DeVellis, 2016, for a comprehensive overview of strategies for developing sound items).

Using a Research Team in the Item Development Process

The initial process of creating items that are intended to measure a latent construct is qualitative in nature, thus, there is utility in incorporating tenants of triangulation of multiple researchers (i.e., a research team, see Carter et al., 2014) from qualitative inquiry into the item development process. Researchers should first individually create a pool of items based on the empirical framework (step 2) and blueprint (step 3). Researchers should seek to develop an exhaustive list of items (i.e., as many as possible) during the first round of item development. The researcher can then edit/reduce their list by looking for redundancy. Once each research team member has created their own list of potential items, they can come together for a series of meetings in which they review and discuss each team member's list of items and eventually come to a consensus about the initial pool of items that will be sent to the expert reviewers (step 5). Conducting a qualitative pilot study with the targeted population is another way that researchers can enhance the rigor in the item development process. Specifically, researchers can conduct individual interviews and/or focus groups with participants that meet the inclusion criteria of the target population. Emergent codes and themes from the qualitative interviews might have utility for guiding the item development process.

Assembling the Instrument

Self-administered questionnaires should be transparent and relatively easy to follow (Fowler, 2014). Researchers are encouraged to implement a standard convention for each element in the measure; for example, using all uppercase letters for the instructions, italicized text for scale points (see Scaling section below), and regular text for test items. Different font styles (e.g., Times New Roman, Calibri) can also be utilized to clearly denote different elements of the survey. Instructions should be as short as possible and include visual cues (e.g., arrows, bolded text) when appropriate, as participants tend to read instructions briefly, if at all (Fowler, 2014). Moreover, definitions should be presented for any vague or abstract terms. For example, The Revised Fit, Stigma, and Value (FSV) Scale, a screening tool for measuring

barriers to seeking personal mental health counseling services, provides respondents with a definition of counseling from the American Counseling Association (Kalkbrenner et al., 2019). The response options or scale points (see Scaling section below) on an instrument tend to appear above the items and can be repeated after every 10 to 15 questions depending on the length of the measure.

Test questions should be as brief and concise as possible and do not necessarily have to be complete sentences. Item stems can have utility for increasing brevity and decreasing respondent fatigue. On The Revised FSV Scale, for example, participants are asked to reply to the following stem: "I am less likely to attend counseling because...." to a number of items (e.g., item "... it would suggest I am unstable," Kalkbrenner et al., 2019, p. 26). Researchers can refer to the theoretical blueprint (step 3) as an aid for ordering the items. When ordering the items, the subject area clusters (intersecting content and domain areas on the blue print) should be interspersed to reduce the likelihood of a response set. Instruments are typically revised and sometimes reassembled throughout steps 5 to 7 as items on the test are usually revised/removed following expert review (step 5) and/or during validity testing (step 7).

Scaling

Researchers should work together to determine the format of measurement or scale for their instrument. Likert scaling is one of the most commonly used scaling formats in the social sciences (DeVellis, 2016). When creating a Likert scale, items are presented in declarative statements with anchor definitions (i.e., response options) that designate fluctuating amounts of agreement or approval of the statement, for example, 1= *strongly disagree*, 2= *disagree*, 3= *neutral*, 4= *agree*, 5= *strongly agree*. It is important to label each anchor definition on the scale. The number and format of anchor definitions should be determined by the construct of measurement (DeVellis, 2016; also see Vagias, 2006, for a variety of Likert scale response anchors). Likert scaling is particularly appropriate for measuring attitudinal constructs (e.g., personality, beliefs, values, or emotions).

Despite the popularity of Likert scales in social science research, a myriad of additional scaling methods are available. Guttman scaling, for example, has usefulness for appraising cumulative/hierarchical constructs in which test takers who endorse a strong statement also endorse milder statements by default (DeVellis, 2016). For example, asking respondents to endorse (select agree or disagree) with each of the following statements, *I feel happy occasionally*, *I feel happy most of the time*, and *I feel happy all of the time*. Someone who selects agree for *I feel happy all of the time* will almost certainly also select agree for *I feel happy most of the time*. Additionally, a binary scale in which respondents are asked to select one of two possible options has particular utility for appraising observed variables with dichotomous response options. For example, asking respondents to indicate (e.g., *yes* or *no*) if they have a high school diploma. Moreover, checklists allow test takers to select multiple response options and are useful when more than one answer might apply to a survey item. For example, a researcher might provide participants with a list of every state in the U.S. and ask them to select all of the states that they have visited.

A semantic differential scale allows one to capture the connotative meaning of stimuli or objects by including unipolar or bipolar adjectives as scale points (DeVellis, 2016). Similarly, on a visual analogue scale respondents are asked to place a mark at a specific point on a line between scale points that represent the opposite ends of a continuum. See DeVellis (2016, pp. 129-130) for examples of semantic differential and visual analogue scales. Finally, a Rasch scale is based on item response theory (Amarnani, 2009) and is centered on the notion that test takers are more likely to respond correctly to items that measure easier degrees of a trait (Boone et al., 2017). Test takers are provided with a more difficult or easier subsequent item based on whether they answered the previous question correctly. Rasch scales have utility for high-stakes testing (e.g., intelligence tests, tests of cognitive ability). Reviewing the intricacies of item response theory and Rasch scaling are beyond the scope of this manuscript, however, refer to Boone et al. (2017) for a primer on Rasch analysis and scaling. Ultimately, researchers should choose their scaling option based on the nature of their construct of measurement (e.g., Likert scaling for attitudinal measures, Rasch scaling for high stakes testing). See DeVellis (2016) for a detailed overview of selecting a scaling option that is consistent with one's construct of measurement.

Step 5: Use Expert Reviewers

Once the raw version of the instrument (initial pool of items and scaling format) is assembled, the measure should be sent to a group of external expert reviewers who are knowledgeable in the content area (Ikart, 2019; Lambie et al., 2017). Experts are sometimes consulted for assistance with item development (step 2), however, different expert reviewers (i.e., people who did not contribute to developing the original pool of items) should be included in this phase to provide a fresh/non-biased perspective. The number of expert reviewers tends to range between three and five, however, upwards of 20 expert reviewers have been noted in the literature (Ikart, 2019). The primary purpose of the expert review process is to maximize the measure's content validity by obtaining feedback from a panel of experts regarding "how relevant they think each item is to what you intend to measure" (DeVellis, 2016, p. 135).

Test developers are responsible for justifying what constitutes an "expert" in a given content area. Expert reviewers (approximately 10+ years' experience) are typically classified into two possible groups for ensuring the rigor and content validity of items, including (a) survey and questionnaire experts, and/or (b) substantive or subject matter experts (Ikart, 2019). Reviewers with survey and questionnaire expertise are well versed in best practices and mechanics of questionnaire design and item development. Subject matter experts have a wealth of knowledge/experience with the construct of measurement and ensure that the collective pool of items sufficiently captures the extensiveness of the construct.

Expert reviewers can be solicited via email list serves associated with professional organizations. Test developers can also use their personal contacts (e.g., current/former professors, employers, co-workers) for suggestions about potential expert reviewers. Expert reviewers can be hired (depending on funding accessibility). Expert reviewers are sometimes added as co-authors of the manuscript if their input significantly influences the measure. In most cases, there is utility in giving the expert reviewers an opportunity to make direct comments on the instrument itself (i.e., track changes in MS Word) as well as soliciting their feedback on a brief survey or form to solicit additional feedback. For example, expert reviewers might be asked to rate on a Likert scale (step 4) the extent to which each survey item represents a content area of the proposed construct of measurement. Researchers also tend to attach an open-response option to Likert scale questions so that reviewers can discuss the reasons behind their ratings. Ikart (2019) provides a comprehensive overview of using expert reviewers in the instrument development process, including but not limited to creating these forms.

Step 6: Recruit Participants

Pilot Testing

Before collecting data from human subjects, researchers should review and obtain proper institutional review board (IRB) approval. Pilot testing (also referred to as preliminary testing) involves administering the instrument to a small developmental sample that is similar to the target population. Pilot testing allows researchers to test their procedures and check for errors in data imputation (e.g., a survey question that asks for a written response, however, the question format is set to only allow a single numeric entry) or technology errors (e.g., particular web browsers that do not support the survey platform or issues with broken or inconsistent hyperlinks). Pilot testing also provides an opportunity to solicit feedback from participants about the content and readability of the items. There are a number of guidelines for what constitutes a small pilot sample, however, pilot samples tend to range between 25 and 150 participants (Browne, 1995; Hertzog, 2008). Pilot study data should be reviewed for information about item content, including clarity and readability as well as for any errors in the administration procedures. Researchers can tentatively compute initial item analyses, for example, inter-item correlations and descriptive statistics. Ideally the pilot sample is 100+ for computing initial item analyses (Field, 2018), however, researchers can compute these analyses with smaller samples as long as they consider the limitations of a small sample size when interpreting the results. Researchers might conduct a factor analysis (step 7) with the pilot data as long as their sample size is sufficient (next section). If pilot study participants highlight issues related to item content and readability, researchers should revise and repeat the pilot process.

Sample Size for the Main Study

Researchers should determine their minimum sample size for the main study prior (a priori) to launching data collection (Mvududu & Sink, 2013). Factor analysis (step 7) is one of the most common statistical tests for validating scores on newly developed measures (Bandalos & Finney, 2019; Benson, 1998; Mvududu & Sink, 2013). In general, larger samples are desirable for factor analysis due to increases in statistical power, however, there is not a clear consensus in the literature for determining the minimal sample size for factor analysis (Knekta et al., 2019). Originally, sample size guidelines for factor analysis were based on general benchmarks. For example, Comrey and Lee (1992) offered the following guidelines for sample size in psychometric research: 50 = *very poor*, 100 = *poor*, 200 = *fair*, 300 = *good*, 500 = *very good*, and ≥ 1,000 = *excellent*. In more recent years, many psychometric researchers determine their minimum a priori sample size by calculating the ratio between the number of participants and the number of estimated parameters or variables being analyzed, sometimes referred to as the subjects-to-variables ratio (STV, Beavers et al., 2013; Mvududu & Sink, 2013). The recommended size of this ratio varies substantially between different psychometricians, from as low as 3:1 to as high as 20:1 (Mvududu & Sink, 2013), however, 10:1 is typically considered acceptable. However, this ratio might be insufficient for estimating the minimum necessary sample size for brief measures (approximately 19 or less items) as the sample size for psychometric studies should include at least 200 participants (Comrey & Lee, 1992).

A number of contemporary psychometricians (e.g., Bandalos & Finney, 2019; Knekta et al., 2019) reject a one size fits all approach for determining sample size. Sample size in psychometric research varies as a function of communality: "amount of variance in the variables that is accounted for by the factor solution, the number of variables per factor, and the interactions of these two conditions" (Bandalos & Finney, 2019, p. 102). Generally, more simplistic models (i.e., fewer items and factors/subscales) require smaller samples; Wolf et al. (2013) demonstrated that a sample size of 30 was sufficient for confirming a unidimensional factor solution with factor loadings > 0.80. Similarly, a sample size as low as 100 can be sufficient for factor analysis with three factors and item communalities that are ≥ 0.70 (Knekta et al., 2019). If, however, communalities are < 0.50, a sample size of 300+ would be required to obtain accurate estimates. Moreover, as the number of factors (subscales) increases the sample size must also increase. For example, a model with seven or more factors would require a sample size of 500+.

Based on the synthesized recommendations of the leading psychometric researchers cited in this section, this writer recommends that test developers determine their a priori minimum sample size by following one of the two following criteria, whichever yields a larger sample: (a) an STV ratio of 10:1 or (b) a sample size of 200 participants. Suppose, for example, a measure is comprised of 50 items. The minimum sample based on an STV ratio of 10:1 would be 10*50 or 500. Before the cessation of data collection, however, researchers should check their sample size with the guidelines provided by Bandalos and Finney (2019) and Knekta et al. (2019, see the previous paragraph) as the unique properties of the data (e.g., communalities, number of items/factors) should be considered when making final decisions about when one has achieved a sufficient sample size.

Obtaining a Sufficient Sample Size: Accessing Participants

There are a variety of strategies for recruiting participants to obtain a sufficient sample size for psychometric analyses (Sharon, 2018). Convenience sampling in public locations (with the proper approvals) can be a cost-effective strategy for accessing participants. When conducting survey research with college students, for example, a researcher might recruit prospective participants as they enter the library or student union. Researchers can also consider using their personal contacts (e.g., current/former professors, employers, co-workers) to distribute recruitment messages for participation in research. For example, a researcher might ask one of their current/former professors to send a recruitment email to all of the students in their department. Researchers who are affiliated with an organization (e.g., university) might have institutional support available to aid in data collection for IRB approved research. As just one example, many universities make their entire student registry (i.e., email addresses of all enrolled students) publicly available. Researchers also sometimes offer small incentives (e.g., small electronic gift cards, bag of candy) to all participants or give participants the option

of entering a raffle to win a prize. When offering incentives for participation in survey research, however, there exist a number of ethical considerations (Singer & Bossarte, 2006). For example, incentives cannot exert undue influence or be coercive, including but not limited to offering excessive monetary compensation. What constitutes an excessive or inappropriate incentive varies by context, thus researchers should work with their research teams, institutional review board, and consult the extant literature to determine an appropriate incentive for a particular study. See Singer and Bossarte (2006) for a detailed overview of practical and ethical considerations when offering incentives in research.

While convenience sampling methods tend to be cost effective, its use comes with a cost to the representativeness of the sample. Data collected via convenience sampling, for example, tends to represent scores from participants who have opportune and a proclivity to participate in survey research (i.e., people who like to take surveys). To this end, more rigorous sampling techniques (e.g., random sampling) tend to enhance the generalizability of results. Alvi (2016) offers a free and comprehensive manual on various sampling techniques in social science research. Depending on funding accessibility, researchers can also hire data collection contracting companies (e.g., IMPAQ, 2020; Qualtrics Sample Services, 2020) for data collection. Qualtrics Sample Services (2020), for example, is a data collection contracting company with a national sampling pool of over 96 million participants and they can recruit random samples, stratified by variables of interest (i.e., adults in U.S. stratified by the most recent census data). Qualtrics Sample Services can also access specific samples, for example, Latinx females in a certain age range, first-generation college students, high school students, and a number of other specific populations. Similarly, Amazon Mechanical Turk (2020) is crowdsourcing marketplace where researchers can recruit prospective participates and offer them monetary compensation to incentivize their voluntary participation. Finally, there are a number of companies (e.g., Redi Data, 2020) that sell randomly generated email lists of a target population.

Electronic Survey Research. Online survey platforms, for example, Qualtrics (2020), REDCap (2020), eSurveysPro (2020), and SurveyMonkey (2020), are becoming increasingly popular. These electronic survey platforms offer user-friendly item construction options (e.g., matrices for building Likert scales, slider options for visual analogue scaling, written response, multiple choice, and more). Most electronic survey platforms generate anonymous electronic links, which can be sent to prospective participants via mass email distribution or posted on websites. In addition, the majority of these platforms also allow users to upload a contact list of prospective participants and use a piped text option to personalize each individual message. Suppose for example, a researcher has a registry spreadsheet of 20,000 prospective participants with their information organized into columns (e.g., first name, last name, email address…). They can personalize the greeting in each message by using a piped text option, which will automatically insert each participant's name in the greeting field (e.g., Dear ${m://FirstName}). Electronic survey platforms also eliminate the need for raw data entry as data are downloaded directly into SPSS or Excel data spreadsheets.

Step 7: Evaluate Validity and Reliability

The final step in initially validating scores on a new measure involves testing for validity (the scale is measuring what it is intended to measure) and reliability (consistency of scores) evidence of the measure and its subscales (Gregory, 2016). In a landmark article, Kane (1992) introduced an argument-based approach to validity based on the notion that making an interpretive argument is "the framework for collecting and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions" (p. 527). According to Kane interpretative arguments can never be proven with absolute certainty. To this end, test developers are tasked with presenting multiple forms of evidence to demonstrate the plausibility of their interpretative argument for validity evidence (Kane, 1992; Kane & Bridgeman, 2017). Validity is a unitary construct, however, there exist a number of sources of validity evidence, including content validity (steps 3 to 5), criterion-related validity, and construct validity (Kane & Bridgeman, 2017; Lenz & Wester, 2017).

Criterion-Related Validity

Demonstrating criterion-related evidence involves examining associations between test scores and a non-test criterion (Kane, 1992). Criterion-related validity evidence includes concurrent validity or the extent to which test scores relate to a non-test criterion in the present. For example, a test developer who compares high school students' scores on an anti-bullying questionnaire to their teachers' ratings of bullying in the classroom is testing criterion-related validity. A high association between the teacher's ratings and anti-bullying scores would yield concurrent validity evidence for the test, as scores on the measure are consistent (concur) with a non-test criterion reference (the teacher's rating). Criterion-related validity evidence can also include predictive validity or the degree to which test scores predict a non-test criterion in the future or past. For example, a test developer might evaluate the predictive validity of a career readiness instrument by testing the extent to which readiness scores predict respondents' future employers' ratings of their job performance. Criterion-related evidence has utility for supporting one's interpretative validity argument, however, demonstrating construct validity evidence is widely considered a cornerstone of validating scores on newly developed tests (Bandalos & Finney, 2019; Benson, 1998).

Construct Validity

Construct validity refers to the extent to which an instrument accurately appraises a theoretical or hypothetical construct and is the most rigorous form of validity evidence for validating scores on newly developed tests (Benson, 1998; Kane & Bridgeman, 2017). Specifically, tests of *internal structure* and *relations with other established theoretical constructs* are two of the most extensively used methods for demonstrating construct validity in social science research (Gregory, 2016; Kane & Bridgeman, 2017; Swank & Mullen, 2017).

Internal Structure and Factor Analysis

Factor analysis, a series of psychometric analysis for testing the dimensionality (internal structure) of the construct of measurement, is probably the most widely used procedure for testing construct validity in social sciences research (Bandalos & Finney, 2019; Benson, 1998; Mvududu & Sink, 2013). There are two primary types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA; Bandalos & Finney, 2019; Mvududu & Sink, 2013).

Exploratory Factor Analysis. The primary purpose of EFA is to uncover the underlying dimensionality within groups of test items by detecting how the items cluster together into subscales (subscales are also known as dimensions or factors), each of which constitute an aspect of the larger construct that the researcher is seeking to measure (Beavers et al., 2013). The EFA is exploratory in nature and the analysis will isolate latent factors that explain the covariance (correlations) among a group of items (Mvududu & Sink, 2013). Prior to computing EFA, the following three preliminary tests should be conducted to test the factorability of the data (i.e., determine if the data set is appropriate for factor analysis): inter-item correlation matrix, Bartlett's test of Sphericity, and the Kaiser–Meyer–Olkin (KMO) Test for Sampling Adequacy (Beavers et al., 2013; Mvududu & Sink, 2013). There are a number of additional important considerations in EFA, including factor extraction, factor rotation, factor retention (Beavers et al., 2013, pp. 4-11), and naming the rotated factors (Mvududu & Sink, 2013, pp. 90 - 91).

Confirmatory Factor Analysis. CFA is a "theory testing strategy" based on structural equation modeling for determining the extent to which the factor solution of an existing measure maintains internal structure with a new sample of participants (Mvududu & Sink, 2013, p. 91). In an instrument development study (or when testing the psychometric properties of an established measure with a new population), researchers should collect data from a new sample and compute a CFA to test the fit between the dimensionality of the hypothesized factor solution with a new sample (Bandalos & Finney, 2019). Model fit is determined by investigating a combination of goodness-of-fit indices such as: incremental, absolute, and parsimonious (Bandalos & Finney, 2019, p. 115). Determining model fit is a complex task and "it is naïve to believe that model fit can be properly assessed by a single index" (Bandalos & Finney, 2019, p. 115). Psychometric researchers offer general cutoff values for particular fix indexes (Hu & Bentler, 1999; Schreiber et al., 2006); however, these values should be used as general guidelines rather than absolute standards. When evaluating model fit, researchers should assess fit

holistically by considering the implications of multiple fit indexes. In addition to evaluating fit indexes, researchers should also consider correlation residuals, parameter estimates, and convergence problems (see Bandalos & Finney, 2019, p. 115) when evaluating model fit.

Relations with Other Established Theoretical Constructs

Examining the relationship between scores on newly developed tests with established theoretical constructs is also a popular method of demonstrating construct validity in social sciences research (Benson, 1998; Strauss & Smith, 2009; Swank & Mullen, 2017). In fact, Benson (1998) refers to testing the relationship between scores on a new test with other theoretically-related measures as "the most crucial" stage in conducting a strong program of construct validation (p. 14). One approach is to test convergent validity or "the relationship among different measures of the same construct" (Strauss & Smith, 2009, p. 1). For example, the developer of a new Depression Severity inventory might test the correlation between scores on their new measure with scores on an established screening tool for depression (e.g., the Beck Depression Inventory). Higher correlations (e.g., $r > 0.5$, see Swank & Mullen, 2017, p. 272) would provide stronger convergent validity evidence, as scores on the new screening tool are similar (converge) with scores on an established measure for appraising the intended construct of measurement (e.g., Depression Severity).

Assessing discriminant validity (also known as divergent validity) is another method of establishing construct validity by demonstrating "that a measure of a construct is unrelated to indicators of theoretically irrelevant constructs in the same domain" (Strauss & Smith, 2009, p. 1). Referring to the example in the previous paragraph, the test developer might correlate scores on their new Depression Severity index with an established measure of Anxiety Severity (e.g., The Beck Anxiety Inventory [BAI]). Based on the extant literature (e.g., Nguyen et al., 2019), one should expect only a minimal-to-moderate relationship between symptoms of anxiety and depression (i.e., divergence between theoretically different contracts in the same domain). Thus, a minimal-to-moderate correlation (e.g., $r < 0.4$, see Swank & Mullen, 2017, p. 272) between scores on the new Depression Severity index and the BAI would support the new scale's discriminant validity as consistent with the extant literature scores suggest that depression and anxiety are separate theoretical constructs (i.e., separate constructs in the same domain).

Employing a Multi-Faceted Approach to Construct Validation

On one level, evaluating construct validity by testing a new measure's relation with other theoretically-related measures presents a potential temporal-validity issue, as one is using an old test to validate scores on a new test (Gregory, 2016). Factor analysis yields information about the internal dimensionality of instrumentation, however, it does not yield evidence about precisely what is being measured (Benson, 1998). To this end, correlating scores on a new test with an established test has greater utility for isolating the precise construct of measurement. It is important to note that no test is inherently valid (i.e., one can only validate scores on a test rather than validate the test itself). Thus, tests are only valid for certain purposes, with particular populations, at specific points of time. Psychometric support for a test is strongest when researchers conduct a series of psychometric studies in which they demonstrate different forms of validity evidence for scores on the test among various populations. To this end, there is utility in employing a multi-faceted method of construct validation. For example, researchers can employ factor analysis to uncover the dimensionality (internal structure) of an instrument as well as testing the convergence/divergence of the measure with other well-established tests. Moreover, initial validity testing can reveal insights for improving the construct and content validity of instrumentation. In such instances, test developers can make revisions to the items and repeat steps 5 to 7. The decision about whether to revise and retest items should be made via research team consensus, which can include consultation with content experts (step 5).

Reliability Evidence

Once a researcher has established validity evidence for scores on their instrument, they should compute a test of the measure's reliability or consistency of scores. There are numerous forms of reliability evidence; test-retest, alternative forms, inter-rater, and internal consistency, (see Bardhoshi & Erford [2017] for a detailed overview of each form of reliability evidence). This author will focus on internal consistency reliability in this manuscript since psychometric researchers tend

to employ cross-sectional research designs, in which data are collected at only one specific point in time.

Cronbach's Coefficient Alpha

Cronbach's coefficient alpha (α) is widely cited (Bardhoshi & Erford, 2017; Cho, 2016; Dunn et al., 2014; McNeish, 2018) as the most commonly used measure of internal consistency reliability across the social sciences and represents the mean value of all possible split-half combinations of the items on a measure or subscale (Cronbach, 1951). Cronbach's coefficient alpha ranges from 0 to 1, with values closer to 1 denoting stronger reliability evidence. There is much debate in the literature regarding the lowest acceptable cutoff value for α. George and Mallery (2003) offer the following guidelines: " $\alpha > .9$ – Excellent, $\alpha > .8$ – Good, $\alpha > .7$ – Acceptable" (p. 231). However, the threshold for an "acceptable" coefficient alpha value should depend on the construct of measurement (Taber, 2018) and the stakes or consequences for test takers that are attached to the test. For example, reliability evidence should be stronger for high-stakes testing (e.g., tests of intelligence or college readiness tests) than for attitudinal screening tools (e.g., interest inventories or non-diagnostic personality tests). Thus, it is the test developer's responsibility to provide a rationale for acceptable internal consistency reliability estimates based on the nature of the test.

Alternatives to Cronbach's Coefficient Alpha

Despite the popularity of Cronbach's coefficient alpha in social sciences research, its use is sometimes called into question (e.g., McNeish, 2018; Taber, 2018). Specifically, coefficient alpha is notoriously misused in instances when the data do not meet certain key assumptions (Dunn et al., 2014) including, the assumption of tau-equivalence or the notion that each scale item equally contributes to the total composite scale score. This is problematic since Cronbach's coefficient alpha tends to underestimate (sometimes substantially) the internal consistency reliability estimate of scores on a scale in the absence of tau-equivalence (McNeish, 2018).

Composite reliability estimates (e.g., McDonald's Omega [ω]) are a viable alternative to Cronbach's coefficient alpha as both estimates produce an internal consistency reliability coefficient based on the ratio between the variance accounted for by each item in relation to the total composite score. McDonald's Omega is advantageous when tau-equivalence is not met as it allows the associations between each item and the total scale to vary. Nájera Catalán (2018) provide a series of recommendations for interpreting ω. McDonald's Omega is the most popular alternative to Cronbach's coefficient alpha (Dunn et al., 2014; McNeish, 2018), however, other options exist including the greatest lower bound (GLB) method and Coefficient *H*. Discussing the intricacies of these estimates is beyond the scope of this manuscript, however, see Bendermacher (2017) for more on the GLB method and McNeish (2018) for more on Coefficient *H*. The overall take-away message is that there is no single, supreme reliability estimate for all tests, as the derivative of each estimate is based on different measurement models. Thus, researchers are tasked with carefully selecting and explaining the most appropriate reliability estimate for their particular study (Cho, 2016; Dunn et al., 2014; McNeish, 2018).

The Measure APPROACH: An Example

Kalkbrenner and Gormley (2020) employed the steps in The MEASURE Approach to develop the Lifestyle Practices and Health Consciousness Inventory (LPHCI). Kalkbrenner and Gormley made their purpose and rationale clear (step 1) by (a) describing their intention to create a measure for appraising lifestyle practices of holistic wellness or integrated dimensions of physical and mental health, (b) exposing a gap in measurement literature for appraising integrated aspects of mental and physical health with a single, relatively brief composite scale, and (c) highlighting the need for such a measure in the integrated primary health care climate in the U.S. Specifically, the LPHCI had great potential for measuring a new latent variable (Global Wellness) for enhancing the future research and practice of practitioners, especially those who work in integrated behavioral health settings.

The empirical framework for the LPHCI (step 2) was developed based on two well-established theoretical models of healthy lifestyle practices for preventing disease and optimizing physical and mental health, including Servan-Schreiber's life-style practices-based anti-cancer method (diet, exercise, stress

management, and avoiding toxins) and Chopra and Fisher's Big Five (mixed nuts, coffee, exercise, vitamin D, and meditation). According to Chopra, Fisher, and Servan-Schreiber lifestyle practices that are only implemented in a single facet of one's life are seldom sufficient for preventing disease; rather, one's engagement in a number of integrated lifestyle practices geared towards enhancing both physical and mental health are essential for promoting their optimal health and wellness (Chopra & Fisher, 2016; Servan-Schreiber, 2009). The theoretical premise of both Servan-Schreiber and Chopra and Fisher's models of holistic wellness was consistent with Kalkbrenner and Gormley (2020)'s aim to develop a screening tool of mental and physical wellness, thus they used these models to set the major theoretical framework for developing a theoretical blueprint (Figure 2) and the initial pool of LPHCI items.

Kalkbrenner and Gormley (2020) created a theoretical blueprint (step 3) for the LPHCI and the content areas (diet, exercise, stress management, and avoiding toxins) were comprised of the four major tenants of Servan-Schreiber's model of mental and physical wellness (Figure 2). The domain areas on the LPHCI blueprint (Figure 2) included frequency, intensity, and duration, which Kalkbrenner and Gormley (2020) adapted from the application-based dimensions of Servan-Schreiber's as well as Chopra and Fisher's models of holistic wellness. On the LPHCI blueprint (Figure 2), for example, the diet, exercise, and avoiding toxins content areas of Servan-Schreiber's model are all related to physical health. Thus, Kalkbrenner and Gormley (2020) included more items in the stress management cells of the blueprint to adjust for the relative importance of holistic wellness (i.e., create a more equal focus on their aim to appraise both mental and physical wellness). The number of items in each intersecting content and domain area on the blueprint (Figure 2) are only approximations of the total number of items that comprised the initial pool of items.

Kalkbrenner and Gormley (2020) began synthesizing content and developing their scale (step 4) by referring to their theoretical framework consisting of Chopra's and Servan-Schreiber's theories as well as a blueprint (Figure 2) to guide the item development process. They made sure the items were brief, clear, and written at approximately a sixth-grade reading level (DeVellis, 2016). For example, LPHCI item 14, "skipped a meal despite feeling hungry" is brief, clear, and written at a fifth-grade Flesh-Kincaid level. Kalkbrenner and Gormley (2020) utilized a research team during the item development process. Each research team member individually developed separate lists of possible LPHCI items based on the empirical framework and blueprint. The team then engaged in a series of meetings until a consensus was reached about the items that became the initial pool of LPHCI items (Kalkbrenner & Gormley 2020).

The initial pool of LPHCI items were sent to three expert reviewers (step 5). Collectively, the reviewers had over 65 years' experience working in medical, academic, and clinical mental health settings. Two of the reviewers were subject matter experts and one was a survey and questionnaire expert. Kalkbrenner and Gormley (2020) then pilot tested the LPHCI with a small sample (N = 125) of the target population; no technology issues emerged and participants did not suggest any revisions to the items, thus researchers proceeded to launch data collection for the main study. Sample size for the main study was based on an STV ratio of 10:1. Specifically, there were a total of 42 LPHCI items to enter into the EFA, thus, the minimum sample based on a STV ratio of 10:1 was 10*42 or 420. Participants were recruited via a data collection service (Qualtrics Sample Services, 2020) to obtain a random national sample (stratified by the U.S. census data) of adults living in the U.S.

Upon the completion of data collection, Kalkbrenner and Gormley (2020) evaluated initial reliability and validity evidence (step 7) for scores on the LPHCI by conducting EFA, CFA, higher-order CFA, and tests of internal consistency reliability. The EFA revealed four latent factors or subscales that comprised the LPHCI. In other words, the EFA identified four groups of observed variables (test items) that clustered together to form four LPHCI subscales. Data from a second sample of participants were entered into a CFA, which revealed acceptable model fit. Finally, tests of internal consistency reliability (Cronbach's coefficient alpha) produced acceptable reliability evidence for the LPHCI scales. Kalkbrenner and Gormley (2020) argued that $\alpha > .70$ was acceptable reliability evidence for the LPHCI because (a) the LPHCI is an attitudinal screening tool, (b) there were no diagnostic or high-stakes implications

for test takers, (c) the construct of measurement was exploratory, and (d) three of the four LPHCI subscales were comprised of relatively few items (shorter scales tend to produce lower reliability estimates). Collectively, the EFA/CFA results and tests of internal consistency reliability produced adequate validity and reliability estimates for the LPHCI. However, a number of poorly worded items were removed during the expert review and validity evidence phases, which made Kalkbrenner and Gormley (2020) concerned about the content validity of the final factor solution. To this end, Kalkbrenner and Gormley (2020) revised the content of the LPHCI items to reflect a more comprehensive scope of the construct of measurement and repeated steps 5 to 7. The results of the second round of item development, data collection, and psychometric testing yielded adequate content validity, construct validity, and internal consistency reliability estimates for scores on the LPHCI.

Conclusions

The MEASURE Approach was designed to provide social scientists with a single resource (one-stop-shop) for outlining seven practical steps in the instrument development process (Table 1). The MEASURE Approach is rooted in classical test theory, with an emphasis on supporting the creation of measures that demonstrate construct validity evidence and evidence of test content (Lenz & Wester, 2017). Such research will require large sample sizes (step 6) and the emergent evidence will be sample-specific until future researchers demonstrate reliability and validity generalizations. A number of further test development considerations can be relevant to social science researchers, for example, cognitive interviews (Peterson et al., 2017), invariance testing (Dimitrov, 2010), higher-order CFA (Credé et al., 2015), using tests outside the normative sample (Hays & Wood, 2017), high-stakes testing (Boone et al., 2017), and cultural/language adaptations (Lenz et al., 2017).

The MEASURE Approach to instrument development has a number of implications for informing the research and the practice of social science professionals. Researchers can refer to The MEASURE Approach to instrument development when evaluating the rigor of existing instrumentation or when creating new measures for use in research. Educators who teach classes in testing, research methods, assessment, or psychometrics can refer to The MEASURE Approach for lesson planning and potentially, include this article as a required or supplemental course reading. Practitioners who work in applied social science fields (e.g., counseling, psychology, or social work) can refer to The MEASURE Approach to instrument development to review the rigor of existing instrumentation before use with their clients. The MEASURE Approach was designed to help social scientists gain a greater understanding of the instrument development process and validating scores on tests, which is consistent with the Standards for Educational and Psychological Testing (2014) and has potential to increase assessment literacy and promote methodological rigor in social sciences research. The overview of the MEASURE Approach presented in this manuscript can serve as a one-stop-shop or a single resource that students and professionals can refer to for outlining empirically supported steps in the instrument development and score validation process.

Table 1. The MEASURE Approach to Instrument Development

Step	Summary Statement
Make the purpose & rationale clear	State the purpose of the instrument development study and provide a rationale for creating a new instrument by (a) reviewing the extant literature and citing any similar instruments that already exist and articulate the construct(s) that the existing measures fail to capture in order to highlight a gap in the existing measurement literature, (b) discuss how the proposed instrument has significant potential to fill the aforementioned gap in the measurement literature, and (c) articulate how filling this gap has potential to advance future research and practice.
Establish empirical framework	Identify a theory (or combination of theories) to set an empirical framework for the item development process. If the literature is lacking an operationalized theory, a researcher can build their own empirical framework by citing a series of empirical sources to provide a rationale for the intended construct of measurement and define the scope of the proposed construct of measurement.
Articulate theoretical blueprint	A theoretical blueprint (Figure 2) is a tool for enhancing the content validity of a measure by organizing the content and domain areas for the construct of measurement and determining the approximate proportion of items that should be developed across each content and domain area.
Synthesize content and scale development	Referring to the empirical framework (step 2) and theoretical blueprint (step 3), researchers should first create a large list of potential items individually. Then, researchers can come together for a meeting(s) to review and compare their separate lists of possible items and negotiate until a consensus is reached about the pool of items that will be sent to the expert reviewers. The initial pool of items should include approximately three to four times as many items that will comprise the final version of the measure.
Use expert reviewers	The initial pool of items is sent to approximately three to five external, expert reviewers. Typically, reviewers are either (a) survey/questionnaire experts, who are well versed in psychometrics and the mechanics of item development or (b) substantive/subject matter experts who are knowledgeable in the content area.
Recruit participants	Administer the instrument to a small pilot sample that is similar to the target population and review the pilot data for data imputation and technology issues as well as participant feedback about item content and readability. Then launch data collection for the main study by following one of the following criteria, whichever yields a larger sample: (a) subjects-to-variables ratio of approximately 10:1 or (b) 200 participants.
Evaluate validity and reliability	Test the validity (the scale is measuring what it is intended to measure) and reliability (consistency of scores) evidence of scores on the measure and its subscales. The MEASURE Approach is centered on demonstrating evidence of construct validity and internal consistency reliability.

References

Alvi, M.H. (2016). *A manual for selecting sampling techniques in research*. Munich Personal RePEc Archive. https://mpra.ub.uni-muenchen.de/70218/1/MPRA_paper_70218.pdf

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. https://www.apa.org/science/programs/testing/standards

Amarnani, R. (2009). Two theories, One theta: A gentle introduction to item response theory as an alternative to classical test theory. *The International Journal of Educational and Psychological Assessment*, *3*, 104–109.

Bandalos, D.L., & Finney, S.J. (2019). Factor analysis: Exploratory and confirmatory. In G.R. Hancock, L. M. Stapleton, & R.O. Mueller (Eds.), The reviewer's guide to quantitative methods in the social sciences (pp. 98-122). Routledge.

Bardhoshi, G., & Erford, B. (2017). Processes and procedures for estimating score reliability and precision. *Measurement and Evaluation in Counseling and Development. 50*(4), 256–263. https://doi.org/10.1080/07481756.2017.1388680

Beavers, A. A., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation*, *18*(5/6), 1-13.

Bendermacher, N. (2017). An unbiased estimator of the greatest lower bound. *Journal of Modern Applied Statistical Methods*, *16*(1), 674–688. https://doi.org/10.22237/jmasm/1493598960

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement, Issues and Practice*, *17*(1), 10–17. https://doi.org/10.1111/j.1745-3992.1998.tb00616.x

Boone, W., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, *4*(1). https://doi.org/10.1080/2331186x.2017.1416898

Browne, R.H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Medicine*, *14*, 1933 – 1940. https://doi.org/10.1002/sim.4780141709

Carter, N., Bryant-Lukosius, D., DiCenso, A., Blythe, J., & Neville, A. (2014). The use of triangulation in qualitative research. *Oncology Nursing Forum*, *41*(5), 545–547. https://doi.org/10.1188/14.ONF.545-547

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651–682. https://doi.org/10.1177/1094428116656239

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Lawrence Erlbaum

Credé, M., & Harms, P. (2015). 25 years of higher-order confirmatory factor analysis in the organizational sciences: A critical review and development of reporting recommendations. *Journal of Organizational Behavior*, *36*(6), 845–872. https://doi.org/10.1002/job.2008

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/BF02310555

DeVellis, R. F. (2016). *Scale development* (4th ed.). Sage.

Dimitrov, D. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. American Counseling Association.

Dimitrov, D. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121–149. https://doi.org/10.1177/0748175610373459

Dunn, T., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *The British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

eSurveysPro. (2020). [Online survey platform software]. Outside Software Inc. https://www.esurveyspro.com/

Fowler, F. J. (2014). *Survey research methods* (5th ed.). Sage

Fu, M., & Zhang, L. (2019). Developing and validating the Career Personality Styles Inventory. *Measurement and Evaluation in Counseling*

and Development, *52*(1), 38–51. https://doi.org/10.1080/07481756.2018.1435193

George, D., & Mallery, P. (2003). *SPSS for windows step by step: A simple guide and reference*. 11.0 update (4th ed.). Allyn & Bacon.

Gregory, R. J. (2016). *Psychological testing: History, principles and applications* (Updated 7th edition). Pearson.

Hays, D., & Wood, C. (2017). Stepping outside the normed sample: Implications for validity. *Measurement and Evaluation in Counseling and Development. 50*(4), 282–288. https://doi.org/10.1080/07481756.2017.1339565

Hertzog, M. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health*, *31*(2), 180–191. https://doi.org/10.1002/nur.20247

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Ikart, E. (2019). Survey questionnaire survey pretesting method: An evaluation of survey questionnaire via expert reviews technique. *Asian Journal of Social Science Studies*, *4*(2). https://doi.org/10.20849/ajsss.v4i2.565

IMPAQ. (2020). [Online survey platform software]. IMPAQ International, LLC. https://impaqint.com/services/survey-research

Kalkbrenner, M.T., & Gormley, B. (2020). Development and initial validation of scores on the Lifestyle Practices and Health Consciousness Inventory (LPHCI). *Measurement and Evaluation in Counseling and Development. 53*(4), 219-237. https://doi.org/10.1080/07481756.2020.1722703

Kalkbrenner, M.T., Neukrug, E.S., & Griffith, S.A., (2019). Appraising counselors' attendance in counseling: The validation and application of the Revised Fit, Stigma, and Value Scale. *Journal of Mental Health Counseling. 4,* 21-35. https://doi.org/10.17744/mehc.41.1.03

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037//0033-2909.112.3.527

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182. https://doi.org/10.1177/0265532209349467

Kane, M., & Bridgeman, B. (2017). *Research on validity theory and practice at ETS*. In R. E. Bennett & M. von Davier (Eds.), *Methodology of educational measurement and assessment. Advancing human assessment: The methodological, psychological and policy contributions of ETS* (p. 489–552). Springer Science + Business Media. https://doi.org/10.1007/978-3-319-58689-2_16

Knekta, E., Runyon, C., & Eddy, S. (2019). One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE Life Sciences Education*, *18*(1). https://doi.org/10.1187/cbe.18-04-0064

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160. https://doi.org/10.1007/BF02288391

Lambie, G., Blount, A., & Mullen, P. (2017). Establishing content-oriented evidence for psychological assessments. *Measurement and Evaluation in Counseling and Development. 50*(4), 210–216. https://doi.org/10.1080/07481756.2017.1336930

Lenz, A., Gómez Soler, I., Dell'Aquilla, J., & Uribe, P. (2017). Translation and cross-cultural adaptation of assessments for use in counseling research. *Measurement and Evaluation in Counseling and Development. 50*(4), 224–231. https://doi.org/10.1080/07481756.2017.1320947

Lenz, A., & Wester, K. (2017). Development and evaluation of assessments for counseling professionals. *Measurement and Evaluation in Counseling and Development. 50*(4), 201–209. https://doi.org/10.1080/07481756.2017.1361303

Maslow, A. (1943). A theory of human motivation. *Psychological Review, 50*(4), 370–396. https://doi.org/10.1037/h0054346

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Menold, J., Jablokow, K., Purzer, S., Ferguson, D., & Ohland, M. (2015). Using an instrument blueprint to support the rigorous development of new surveys and assessments in engineering education. *ASEE Annual Conference and Exposition, Conference Proceedings*, *122*. American Society for Engineering Education.

Mvududu, N. H., & Sink, C. A. (2013). Factor analysis in counseling research and practice. *Counseling*

Outcome Research and Evaluation, 4(2), 75-98. https://doi.org/10.1177/2150137813494766

Nájera Catalán, H. (2018). Reliability, population classification and weighting in multidimensional poverty measurement: A Monte Carlo Study. *Social Indicators Research*, *142*(3), 887–910. https://doi.org/10.1007/s11205-018-1950-z

Nguyen, D., Wright, E., Dedding, C., Pham, T., & Bunders, J. (2019). Low self-esteem and its association with anxiety, depression, and suicidal ideation in Vietnamese secondary school students: A cross-sectional study. *Frontiers in Psychiatry*, *10*(SEP). https://doi.org/10.3389/fpsyt.2019.00698

Peterson, C., Peterson, N., & Powell, K. (2017). Cognitive interviewing for item development: Validity evidence based on content and response Processes. *Measurement and Evaluation in Counseling and Development. 50*(4), 217–223. https://doi.org/10.1080/07481756.2017.1339564

Qualtrics. (2020). [Online survey platform software]. SAP. https://www.qualtrics.com/

Qualtrics Sample Services [Online sampling service service]. (2020). https://www.qualtrics.com/research-services/online-sample/

REDCap. (2020). [Online survey platform software]. Vanderbilt. https://www.project-redcap.org/

Redi Data. (2020). [data services and direct marketing solutions] http://www.redidata.com/

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *Journal of Educational Research*, *99*(6), 323-337. https://doi:10.3200/JOER.99.6.323-338

Sharon, T. (2018). *43 ways to find participants for research.* Medium. https://medium.com/@tsharon/43-ways-to-find-participants-for-research-ba4ddcc2255b

Singer, E., & Bossarte, R. M. (2006). Incentives for Survey Participation: When are they "Coercive"? *American Journal of Preventive Medicine*, *31*(5), 411–418. https://doi.org/10.1016/j.amepre.2006.07.013

Strauss, M., & Smith, G. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*(1), 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Summers, S., & Summers, S. (1992). Instrument development: Writing the items. *Journal of Post Anesthesia Nursing*, *7*(6), 407–410. http://search.proquest.com/docview/73394064/

SurveyMonkey. (2020). [Online survey platform software]. Bain & Company, Inc., Fred Reichheld and Satmetrix Systems, Inc. https://www.surveymonkey.com/?ut_source=sem_lp&ut_source2=sem&ut_source3=header

Swank, J., & Mullen, P. (2017). Evaluating evidence for conceptually related constructs using bivariate correlations. *Measurement and Evaluation in Counseling and Development. 50*(4), 270–274. https://doi.org/10.1080/07481756.2017.1339562

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tate, K., Bloom, M., Tassara, M., & Caperton, W. (2014). Counselor competence, performance assessment, and program evaluation: Using psychometric instruments. *Measurement and Evaluation in Counseling and Development*, *47*(4), 291–306. https://doi.org/10.1177/0748175614538063

Urdan, T. C. (2010). *Statistics in Plain English* (3rd ed.). New York, NY: Routledge.

Vagias, W. M. (2006). *Likert-type scale response anchors.* https://www.uc.edu/content/dam/uc/sas/docs/Assessment/likert-type%20response%20anchors.pdf

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Citation:
Kalkbrenner, M. (2021). A Practical Guide to Instrument Development and Score Validation in Social Sciences Research: The MEASURE Approach. *Practical Assessment, Research & Evaluation*, 26(1). Available online: https://scholarworks.umass.edu/pare/vol26/iss1/1/

Corresponding Author

Michael Kalkbrenner
Office: O'Donnell 202 J
New Mexico State University
Las Cruces, NM, 88003-8001

Email: mkalk001 [at] nmsu.edu