Establishing Validity and Cross-Context Equivalence of Measures and Indicators

Edward A. Frongillo, PhD; Tom Baranowski, PhD; Amy F. Subar, PhD, RD; Janet A. Tooze, PhD; Sharon I. Kirkpatrick, PhD, RD

ARTICLE INFORMATION

Article history:
Submitted 29 March 2018
Accepted 6 September 2018
Available online 20 November 2018

Keywords:
Validity
Reliability
Accuracy
Measures
Equivalence

2212-2672/Copyright © 2019 by the Academy of Nutrition and Dietetics.
https://doi.org/10.1016/j.jand.2018.09.005

ABSTRACT

Quantitative research depends on using measures to collect data that are valid (ie, reflect well the phenomena of interest) and perform equivalently across contexts. Demonstrating validity and cross-context equivalence requires specifically designed studies, but many such studies have problems that have limited their usefulness. This article explains validity and cross-context equivalence of measures (and important related concepts) and clarifies how to establish them. Validation is the process of determining whether a measure or indicator is suitable for providing useful analytical measurement for a given purpose and context. Cross-context equivalence means that a measure performs comparably across contexts. Four types of equivalence are construct, item, measurement, and scalar. Establishing validity and cross-context equivalence requires representing mathematically the errors (ie, imprecision, undependability, and inaccuracy) of a measure and using appropriate statistical methods to quantify these errors. Studies aiming to provide evidence about the validity of a measure need to clarify the purpose and context for use of that measure. Choose one of the two conceptual systems for validation; obtain data to establish the extent to which the measure is well constructed, reliable, and accurate; and use analytic methods beyond simple correlations to provide a basis for making reasoned judgment about whether the measure provides useful analytic measurement for the particular purpose(s) and context. Establishing accuracy of a measure requires having available other measures known to be accurate as comparators; in the case that no other measure understood to be more accurate is available, then the study will be able to establish agreement rather than validity.
J Acad Nutr Diet. 2019;119(11):1817-1830.

QUANTITATIVE RESEARCH DEPENDS ON THE USE OF measures to collect data that reflect well the phenomena of interest; that is, that are valid. Demonstrating that measures are valid requires studies that are specifically designed to determine the extent to which measures reflect the phenomena of interest. The process of validation is essential in selecting measures appropriate for quantitative research, understanding the current limitations in measures, and determining where further research or development is needed. Five problems are common in current validation studies. First, there is lack of clarity about terms used to describe and analyze validation studies. Second, this lack of clarity is partly because two different conceptual systems of validation are used in the field of nutrition science. Third, the difference between establishing validity of a measure and establishing agreement between two measures is not always recognized, leading to confusion and inconsistent interpretation across validation studies. Fourth, proper procedures and principles of validation are not always followed, leading to some purported validations that do not actually validate measures. Fifth, the specific applications of the validated measures are often not specified; that is, the context and purpose for which the measure is appropriate are not always stated, and it is often assumed that validation for one purpose and in one context extends to another. Partly as a consequence of these problems, attempts to develop better measures are not as rigorous as they could be, limiting what can be learned from them. Furthermore, for many studies or topics, not having measures that are equivalent across contexts (ie, locations or settings) makes comparisons difficult. This article explains the role of validity and cross-context equivalence of measures and important related concepts and clarifies principles and methods for establishing validity and equivalence across contexts.

The Continuing Professional Education (CPE) quiz for this article is available for free to Academy members through the MyCDRGo app (available for iOS and Android devices) and through www.jandonline.org (click on "CPE" in the menu and then "Academy Journal CPE Articles"). Log in with your Academy of Nutrition and Dietetics or Commission on Dietetic Registration username and password, click "Journal Article Quiz" on the next page, then click the "Additional Journal CPE quizzes" button to view a list of available quizzes. Non-members may take CPE quizzes by sending a request to journal@eatright.org. There is a fee of $45 per quiz (includes quiz and copy of article) for non-member Journal CPE. CPE quizzes are valid for 1 year after the issue date in which the articles are published.

CONCEPTS AND ROLE OF VALIDITY AND EQUIVALENCE ACROSS CONTEXTS

Measures and Indicators

A goal of quantitative research is to obtain data that can be put to use by people in a position to make decisions. Data generally are obtained and put to use through a series of six steps: a method (eg, a 24-hour recall of dietary intake) employs a measurement process that yields a measure (eg, fat intake in grams per day) that produces a measurement (ie, a value of fat intake for a given person) that is used for a particular purpose and context to make an inference, conclusion, or result that leads to a decision. Although one can think about validity as applicable to any of these six steps (eg, validity of a method or validity of an inference), this article focuses on the validity and cross-context equivalence of measures and indicators.

Measures assign numbers to people or things to represent the relative amounts of a property.[1] For example, height is a property of a child that can be measured, assigning a number to a child. Multiple properties of an individual may be measured that may differ by whether they are objective or primarily or fully subjective, fundamentally unidimensional or multidimensional, or vary over time (Figure 1).

Indicators are different from measures. Indicators reflect the presence or absence of a given property. In social and epidemiologic research on human beings in particular, indicators are important to describe populations. Indicators are constructed by either defining classes directly (eg, race and ethnicity) or classifying values of a single measure or an index or scale calculated from multiple measures, with the classification based on degree or specific meaning.[2] For example, the height of a child is a measure, whereas whether or not a child is stunted (ie, height below −2 standard deviations of the sex- and age-appropriate growth standard) is an indicator. For an indicator to be useful, an understanding of which values of a measure, index, or scale are considered good (vs poor) is needed.[3] Usefulness of an indicator also depends on its relevance to purpose and context, absolute and relative cost, conceptual and empirical comparability (ie, equivalence), sensitivity to real differences and changes in time, and credibility (ie, acceptance by those who will use the indicator).[4]

Measures (and indicators) can be objective or subjective. Objective means that the measure is external to the mind, whereas subjective means that it emanates from the mind, reflecting the thoughts (eg, memories and beliefs) and/or feelings of a respondent. Differentiating these is not always straightforward. For example, food insecurity may have a subjective component (eg, how one experiences lack of food), yet there is also an objective reality (eg, no food in the home or a skipped meal). In contrast, psychologic traits, attitudes, or moods are inherently subjective. For example, it may be possible to observe behaviors that indicate depression, but without report of depressed mood or loss of interest or pleasure along with other criteria, which is subjective, a person would not be considered to be depressed. Some other behaviors or physical or social properties (eg, income) could be measured objectively without the individual's involvement (eg, by examining financial records), but are often self-reported, which could have errors of memory, numeracy, or even self-presentation manipulation. Furthermore, often a property is assumed to be unidimensional (ie, have one underlying construct), but some properties are inherently multidimensional; eg, dietary intake can represent consumption of a complex array of foods and nutrients as well as eating patterns and dietary behaviors. Also, some properties, such as a blood type, are not expected to change over time, but others, such as dietary intakes, often do vary over time.

In nutrition sciences, both objective and subjective measures (and indicators) are used. Physical examinations, for example, can produce objective measures such as weight, height, and body composition that can be observed. In contrast, self-report dietary assessment instruments such as 24-hour recalls, food records, and food frequency questionnaires as well as household food insecurity questionnaires produce measures that are subjective or include substantial subjectivity. These measures result from the experience of respondents as conveyed through psychological processes of cognition (ie, memory and perception) and affect, and are influenced by motivation. For assessment of dietary intake over a 24-hour period through recall, for example, the dietary intake that occurred during the 24-hour period theoretically could have been objectively observed, but the recall of the intake is subjective. One purpose of self-reported dietary assessment is to measure usual intake, a construct that is often latent; the true usual intake that exists is difficult to observe because intake generally varies day to day. The experience of household food insecurity has been shown in ethnographic studies to have multiple latent constructs (eg, feelings of uncertainty and deprivation, social acceptability of food access, quantity of food, and quality of food), but commonly used experience-based measures of household or individual food insecurity have been built assuming one latent construct, not including items that would measure each of the constructs (particularly deprivation and social unacceptability).[5]

RESEARCH SNAPSHOT

Research Question: What principles and methods should be used for establishing validity of measures and indicators and their equivalence across contexts?

Key Findings: Studies aiming to provide evidence about the validity of a measure need to clarify the purpose and context for use of that measure; choose one of the two conceptual systems for validation; obtain data to establish the extent to which the measure is well constructed, reliable, and accurate; and use analytic methods beyond simple correlations to provide a basis for making reasoned judgment about whether the measure provides useful analytic measurement for the particular purpose(s) and context.

Validity

Validation is the process of determining whether a measure or indicator is suitable for providing useful analytical measurement[6] for a given purpose and context.[7] A measure or indicator is valid in the case that each of six criteria are met: its construction is well grounded in theory; its performance is consistent with that theory; it is precise, dependable, and accurate[7] within specified performance standards; and its

Type	Definition	Example	Subjective?	Multidimensional?	Varies over time?
Trait	Fundamental characteristic that describes someone's personality, cannot be changed	Extraversion	Yes	No, not each trait, but personality as a whole is	No
Attitude	Opinions; organization of beliefs, feelings, and behavioral tendencies that have affective, behavioral, and cognitive components	Opinion about taste of foods	Yes	Yes: belief, feelings, behavioral tendencies	Yes
State	Moods	Depression	Yes	No, not each mood	Yes
Social Standing	A person's social standing or class	Food security	No	Yes	Yes
Knowledge	Information or awareness acquired by an individual	Reading level	No	Possibly	Yes
Behavior	Way in which one acts or interacts	Eating	No, but reporting of it typically is	Yes, because they are inherently linked with time, but not at a particular time, if we think about "usual" as well	Yes
Physical or biological property	Sensory or biological qualities	Height	No	Each is usually unidimensional	Some do (height), some do not (sex)

Figure 1. Properties of individuals that may be measured.

accuracy is attributable to the well-grounded theory for that purpose and context.[8] That is, a valid measure or indicator will be well constructed and perform according to its construction; reliable (ie, precise and dependable); and accurate, with accuracy that is attributable to the theory underlying the construction.[3] For example, the measure of household food insecurity used in the United States has been shown to be valid for estimating prevalence and differentiating households because construction of the measure is well grounded in understanding of food insecurity from ethnographic studies and its performance is consistent with that understanding; the set of items used in the measure are reliable (ie, internally consistent); and the measure accurately estimates prevalence of groups and differentiates households, with the accuracy attributable to the well-grounded understanding.[8]

Regarding the latter criteria, the terms precision, dependability, and accuracy are sometimes confused in the literature. Precision, also called technical error of measurement, is the extent to which repeated measurements yield the same value. Precision is achieved by careful measurement and redundancy.[8] Imprecision is usually considered to be random error, and is estimated by conducting a test–retest exercise or calculating internal consistency of a set of questionnaire items.[8] Precision may vary with the person being measured (eg, some children fidget more than others) and the person doing the measuring (eg, some study staff are more consistent measurers than others).

Dependability, in contrast to precision, is the extent to which differences in a measure consistently reflect actual differences in the property.[8] For example, in dietary assessment using a 24-hour recall to measure usual intake, day-to-day variation in dietary intake is a source of undependability. Undependability is usually considered random error, although it can encompass systematic error, such as the progressive shortening of adult height during the day because of disc compression or progressive lower reporting of intake across days in diet diaries. Dependability is achieved by understanding and avoiding the sources of random or systematic error that are threats to it, and is estimated by conducting a test–retest exercise with a longer time between measurements than for estimating imprecision.[8] The time between measurements has to be chosen carefully when natural or intervention-induced changes may occur over time. Dependability may vary with the person being measured, as has been shown for infant length.[9] Taken together, precision and dependability constitute reliability.

Accuracy is the extent to which a measure provides unbiased assessment of a property; that is, without systematic error (ie, bias).[8] Accuracy is achieved by construction of the measure, which is based on well-grounded understanding of the property. The term *well grounded* means that there is a strong theoretical basis for why the measure is a reasonable way of measuring the property of interest.[8] Accuracy can be demonstrated by in-depth analysis and by relating the measure to a criterion measure, which ideally would be a more definitive measure; sometimes for demonstrating accuracy at a group level, a determinant or consequence can be used as a criterion measure. Demonstrating that accuracy is actually due to the well-grounded understanding is important. Otherwise, a measure that is only apparently accurate may be useless. For example, the apparent accuracy of a measure of lean body mass of small animals using total body electrical conductivity was attributable to the measure's relation to total body weight and not to lean body mass as the developers of the instrument had assumed.[10] Therefore, attribution of accuracy can only be achieved by careful study design and availability of other measures and information, and it is demonstrated by comparison with competing measures and examination of alternative explanations.[8]

Measures can be categorized according to a hierarchy of accuracy: definitive, reference, and field (ie, routine).[11] A definitive measure, sometimes called a gold-standard measure, relies on first principles (ie, the fundamental and self-evident basis) to achieve high accuracy; that is, with little or no error, and it reflects in a fundamental way the theoretical structure of the property it purports to represent. A reference measure directly and closely relates to the property of interest, but typically does not reflect the fundamental theoretical structure of the property as closely as does a definitive measure. Ideally the accuracy of a reference measure is demonstrated by comparison to a definitive measure.[11] A field measure is usually fast, routine, and inexpensive, requiring relatively unsophisticated personnel and technology. Accuracy of a field measure is often best demonstrated by comparison to a reference measure (or definitive measure when available).[11] For example, in assessment of body composition, specifically the property of the percentage of the body that is fat, definitive measures of the composition of body tissues have come from cadaver studies.[12] Reference measures include densitometry, potassium-40 counting, and dual-energy x-ray absorptiometry. Field measures include anthropometry (eg, weight, height, and skinfold thicknesses), total body electrical conductivity, and bioelectrical impedance.

Identifying definitive measures against which to establish the accuracy of reference measures is a particular challenge in dietary assessment. An example is a protocol in which researchers observe what children eat at school on 1 day (ie, a definitive measure) and then compare that with the children's recall of what they ate on that day (ie, a reference or field measure). This comparison can be done for each food as well as the child's total diet.[13] A body of research has been conducted using recovery biomarkers as reference measures of dietary intake, including doubly-labeled water for energy intake, 24-hour urinary nitrogen for protein intake, and 24-hour urinary potassium and sodium for intakes of these elements.[14] Doubly-labeled water and urinary biomarkers might be definitive measures for the 7 to 14 days and 1 day, respectively, that were specifically measured, but would be a reference measure when used to assess intake for longer periods.

Level	Possible Purposes
Groups of households or people	• Estimation of prevalence (What is the magnitude of the problem?) • Determination of causes and consequences (Why are they affected and what are effects?) • Early warning (When is action needed?) • Targeting (Who will receive which action?) • Monitoring (How is the situation changing?) • Impact evaluation of programs (Has the action made a difference?)
Individual households or people	• Screening (Is the household or individual at risk?) • Diagnosis of problem (Does the household or individual have the problem, and what are the salient causes?) • Diagnosis of solution (What is the most appropriate action?) • Monitoring (How is the situation changing?)

Figure 2. Possible purposes of measures and indicators at group and individual levels.[2,3]

In addition to providing evidence that the six criteria for construction, reliability, and accuracy have been met, a claim that a measure is valid must always include a statement about the purpose and context of the measure because it can be valid for one purpose (eg, estimating prevalence) but not others (eg, estimating response to an intervention) and in one context (eg, old adults in cities) but not others (eg, young adults in rural areas). That is, validity is not inherent to a measure; it refers to the suitability of a measure when used for a particular purpose in a particular context.[8] Possible purposes can either be at the group or individual level (Figure 2).[2,3] For example, a single 24-hour recall taken on a group of people may produce a valid estimate of mean usual intake of food and nutrients of a group because random errors tend to cancel out, but not be useful in estimating the usual intake of an individual because of random errors due to imperfect memory (ie, imprecision) and day-to-day variation (ie, undependability).

A conclusion about the validity of a measure is based on judgment, in that validity is established by comparing the performance of the measure being assessed to preestablished performance standards. The word validation is sometimes used erroneously to mean either that something has been established as truth even when it has not or that the measure reflects the way that the world actually is. Validation does not necessarily establish truth; instead, validation establishes legitimacy,[15] although the word true is often used to mean the measurement that would be obtained if using a perfect measure.

Studies of agreement, reliability, and calibration are commonly mislabeled as validations. Studies of agreement compare one measure to another in cases in which neither measure is expected to be more accurate than the other for the property of interest; for example, usual intake measured with a food frequency questionnaire compared with several 24-hour recalls.[16] Agreement studies (also known as relative validations) are common in dietary assessment, given that measuring true usual intake over a prolonged period of time is either difficult or impossible to do. Studies of reliability can only demonstrate that a measure gives similar results on two different occasions; given that the true value may have changed between occasions, reliability studies cannot determine validity. Studies of calibration bring one set of values into line with another, generally by use of a regression equation.[17] Total body electrical conductivity, in which a quantity reported by the machine is used to predict lean body mass, provides an example of calibration.[10] Furthermore, food records or 24-hour dietary recalls for multiple days have been used to calibrate dietary intake from food frequency questionnaires.[17]

COMPARISON OF THE BIOMETRIC AND PSYCHOMETRIC CONCEPTUAL SYSTEMS FOR VALIDITY

The above definitions and explanations of validity come from a biometric conceptual system used in the chemical and biological sciences in which objective measures can be obtained.[5] Another conceptual system to understand validity of measures comes from psychometrics.[16,18,19] Measuring social and behavioral properties has been influenced by psychometrics; many social and behavioral assessments depend on self-report, are subjective, and are not observable,[16] although there are exceptions (eg, accelerometers for physical activity). Both conceptual systems are used in nutrition sciences, which brings together the chemical and biological sciences with the social and behavioral sciences. For example, validation of measures of household food insecurity often has been conducted using the biometric conceptual system, whereas validation of measures of dietary intake has most often been conducted using the psychometric system.

The two conceptual systems, although different in some ways, can be mapped to each other (Figure 3). The definition of validity in the biometric conceptual system focuses on whether measures can provide useful measurement for a given purpose and context. The six criteria of validity refer to the features of measures needed to achieve validity. Validity in the psychometric conceptual system also focuses on whether a measure "is useful scientifically"[17] and "if it does what it is intended to do"[18]; another definition is the extent to which a measure yields the true scientific value free of errors, including bias; that is, inaccuracy.[19] In the psychometric system, there are several types of validity (see Figure 3) that correspond to the major functions of psychological measures[16,18] and represent ways to demonstrate validity. "Reliability is a necessary but not sufficient condition for validity."[18]

Equivalence Across Contexts

Equivalence here means that a measure performs consistently across contexts.[21,22] Four types of cross-context equivalence can be defined: construct, item, measurement, and scalar.[3] Construct equivalence means that "the same construct is measured across contexts, even if the measures used are not identical"[3]; in this case, the constructs measured are comparable. For example, the construct of food management strategies in response to household or individual food insecurity is equivalent across contexts, but the items needed to measure this construct will differ markedly across contexts because the specific management strategies available and used depend on what is possible in the context. Item equivalence means that the "same construct is measured across contexts and the content of each item used is perceived and interpreted in the same way across contexts"[3]; in this case, the same items used across contexts mean, and are interpreted, the same. For example, the nine items used to assess household food insecurity in the Household Food Insecurity Access Scale were developed to be item equivalent across contexts.[23] Measurement equivalence means that the "constructs, items, and units are the same across contexts (ie, the difference in scores between two individuals means the same across contexts)"[3]; in this case, the order of households or individuals is comparable across contexts. Scalar equivalence means the "same as measurement equivalence, but in addition the definition of zero is the same across contexts"[3]; in this case, average scores and prevalence values are comparable. For example, only three of the nine items in the Household Food Insecurity Access Scale that assess hunger are scalar equivalent.[24] Scalar equivalence is ideal because scores can be meaningfully added and subtracted, and the scores of a given measure and prevalence of an indicator across contexts are directly comparable. Scalar equivalence is often difficult to achieve, and measurement equivalence, when possible to achieve, provides for meaningful comparisons across contexts for many purposes.

In the psychometrics literature, a parallel concept to equivalence is measurement invariance; that is, how the psychometric characteristics of a measure are preserved across contexts.[25,26] A measure is invariant when individuals from different contexts with the same status for a construct receive the same score on the measure.[27] Confirmatory factor analysis is often used to examine the construct validity of a set of questionnaire items, and it can be used to provide evidence of factorial invariance, and thus measurement invariance, across contexts. Five forms of factor invariance have been described.[25] Dimensional invariance addresses whether the same number of factors is present in each context. Configural invariance addresses whether factors are associated with the same items in each context. Metric invariance addresses whether the factors have the same meanings in each context. Strong factorial invariance addresses whether group means can be compared meaningfully across contexts. Strict factorial invariance addresses whether

	Biometric Conceptual System		Psychometric Conceptual System	
Category	**Criteria**	**Definition**	**Concept**	**Definition**
Validity	—	Whether a measure is suitable for providing useful analytical measurement for a given purpose and context.[3,8]	—	Whether a measure is useful scientifically and does what it is intended to do.[18] Extent to which a measurement is representative of the true scientific value, taking true to mean an exact representation of what happened, free from all possible sources of error, including bias.[19]
Construction	Well-constructed	Grounded in an understanding of the underlying phenomenon being measured.	Face validity Content validity	Extent to which a measure looks like it will, or appears to, provide the desired information. Extent to which a measure covers all aspects of the intended behavioral or physiological domains or dimensions.
	Consistent Performance	Performance is consistent with that understanding.	Convergent (or construct) validity	Extent of the agreement with another (noncriterion) measure that should assess the same parameter based on face and content validity.
Reliability	Precision	Yield the same value upon repetition of the measurement when the property does not differ	Reliability	Extent to which test scores are consistent from 1 test administration to the next; keeping as many conditions as possible unchanged. Extent to which test scores are consistent when measurements are taken by different people (or instruments) using the same methods or at different times by the same person.
	Dependability	Variability or differences in the measurement consistently reflect variability or differences in the property being measured.	Behavioral reliability	Extent to which stability in behavior has been considered when assessing other aspects of reliability.
Accuracy	Accuracy	Extent to which a measure provides unbiased assessment of the property in comparison to a more definitive measure	Criterion validity	Extent of the agreement between a measure and another already held as being a criterion or gold standard.
	Attribution of accuracy	Accuracy is attributable to the well-grounded understanding for that purpose and context.	Discriminant validity	Extent to which the measure is novel and does not simply reflect some other variable.[20]

Figure 3. Comparison of two systems for conceptualizing validity from the chemical and biological sciences and psychometrics.

group means and variances can be compared meaningfully across contexts. The concepts of dimensional and metric invariance are related to that of construct equivalence. The concept of configural invariance is related to that of item equivalence. The concepts of strong and strict invariance are related to that of scalar equivalence.

Errors in Measures

Establishing validity and equivalence across contexts requires the collection of data using measures and indicators, mathematical representation of the quantities needed from the data, and analysis using appropriate statistical methods. Among the quantities needed are estimates of errors in measures; that is, imprecision, undependability, and inaccuracy. The biometric conceptual system corresponds to a standard mathematical and statistical representation of errors.[28-30] If X_{ij} is a measurement of individual i at occasion j and T_i is the true value of that individual, then for a given measurement,

$$X_{ij} = T_i + E_{ij} \tag{1}$$

where E_{ij} is the measurement error. If $X_{ij} \neq T_i$ then $E_{ij} \neq 0$. The error E_{ij} can be non-zero because of inaccuracy or unreliability or both, and for a single measurement X_{ij} it is impossible to distinguish between these two sources of the error. Only by having repeated measurements on multiple individuals can inaccuracy and unreliability be estimated and differentiated. Alternatively, equation (1) can be written:

$$E_{ij} = X_{ij} - T_i \tag{2}$$

The square of equation (2) yields the mean square error (ie, the total error), which summed across individuals is:[28]

$$\begin{aligned}&\sum (X_{ij} - T_i)^2 \Big/ (n-1)\\ &\quad = \sum (X_{i\blacksquare} - T_i)^2 \Big/ (n-1) + \sum (X_{ij} - X_{i\blacksquare})^2 \Big/ (n-1)\end{aligned} \tag{3}$$

where $X_{i\blacksquare}$ is the mean for individual i across occasions. That is, the mean square error is the sum of the square of the bias (ie, inaccuracy) and the variance of unreliability. The variance of unreliability can be further partitioned into variances of undependability and imprecision, and the imprecision into variances for inter- and intrarater components. Equation (3) illustrates that there can be trade-offs in mean square error between inaccuracy and unreliability such that, of two measures with the same mean square error, one could be highly accurate but unreliable and the other inaccurate but highly reliable. In some situations, for example, a measure with small bias that is highly reliable may be more useful than one with no bias that is highly unreliable. The sum of inaccuracy, undependability, and imprecision, which is quantified as the mean square error, is the total error.[7]

This representation of errors assumes that a true value is plausible for a given individual, such as for physical properties, demographic characteristics, or behaviors.[30] In contrast, measures of psychological states, attitudes, or knowledge do not have a true value that is plausible for an individual. Psychometricians instead assume that there exists a distribution of measurements for an individual, with the mean of the measurements substituting for the true value and representing a latent construct for the individual.[30] This distinction in part has led to the development of the psychometric conceptual system for validity. Psychometrics makes use of two theories that assume latent constructs, Classical Test Theory and Item Response Theory, as a basis for the development and testing of measures formed by scaling questionnaire items. Classical Test Theory models the observed manifestation of constructs through items as a function of a true score and a random error as presented in the previous section.[31] Factor analysis is often used to examine the relationship between constructs and items. Item Response Theory models the relationship between a latent construct and the probability of particular responses to items using a nonlinear monotonic function that bounds the probabilities to zero and one. The simplest model defines a difficulty or severity parameter that differentiates items. Item Response Theory provides statistics to evaluate the fit of individual items. Both theories assume that the probability of affirming an item is monotonically related to status on the latent construct and that status on the latent construct accounts fully for responses to items in the scale.[31]

This section has discussed representation of absolute errors, which means that errors are quantified as differences or squares of differences, with the differences being in the units of the measure. An alternative representation is relative error, which expresses differences relative to (ie, divided by) estimates of the variation in the sample being studied. For example, a coefficient of unreliability estimates relative error using a correlation.

ESTABLISHING VALIDITY AND CROSS-CONTEXT EQUIVALENCE

Establishing validity is inherently an example of a prediction problem in that it is desired to know how well measure X_{ij} predicts the true property T_i (as represented by a definitive or reference measure). In any prediction problem, including establishing validity, quantifying the absolute error in the units of the measure is more important, informative, and useful than quantifying the relative error.[32] That is, confidence or prediction intervals that represent the error in predicting T_i from X_{ij} in absolute units are needed. For example, when examining the validity of skinfold thickness as a measure of body fat mass in adults, it is important to know the error in units of body fat mass (ie, kilograms) rather than just the relative error (eg, the correlation between skinfold thickness and body fat mass). There are two interrelated reasons for this. First, the absolute error quantifies in meaningful units how close the measure comes on average to the property of interest, allowing judgment of whether the closeness is sufficient to be useful. Second, the relative error is confounded by the underlying variation in the sample, and therefore cannot be easily applied to other samples in which the underlying variation may be different.

For these reasons, regression methods are preferred to simple correlation methods for quantifying validity.[33-35] A bivariate regression model estimates the correlation coefficient but goes beyond that to capture comprehensively the relationship between the measure being tested and the definitive or reference measure. The correspondence between the two measures can be assessed by the regression of the definitive or reference measure on the test measure. If the units are the same, then a test measure that is accurate and reliable could result in an intercept of 0, a slope of 1, and points falling close to the predicted straight line. A standard residual plot can help reveal the features of the relationship. Even if the intercept is not 0, the slope is not 1, and/or the relationship does not correspond to a straight line, the test measure may be accurate and reliable within the range of data values observed, with calibration perhaps required to convert the values from the test measure to meaningful values, as is done with total body electrical conductivity. Regression analysis is also preferred because the regression model can be used to examine attribution of accuracy by adding measures of potential alternative explanations to determine whether the measure being tested predicts the definite or reference measures beyond what can be predicted by the alternative measures.

For indicators such as whether or not a young child has faltered in growth[36] or whether or not a child has experienced food insecurity,[37] T_i is binary. For example, children with growth faltering have T_i positive and children without growth faltering have T_i negative. When T_i is binary, sensitivity—specificity analysis often is used to assess validity.[16] In this analysis, sensitivity (ie, proportion of those with T_i positive that are correctly classified as positive by measure X_{ij}) and specificity (ie, proportion of those with T_i negative that are correctly classified as negative by measure X_{ij}) are calculated for a continuous measure X_{ij} at all possible cut-points of the observed distribution of X_{ij}. Then, the area under the receiver operating characteristic curve (ie, the plot of sensitivity vs 1-sensitivity) is used to quantify accuracy[38]; the area under the curve theoretically ranges from 0.5 (chance) to 1.0 (perfect accuracy). For the example of the accuracy of a measure of growth faltering that could be used in epidemiologic studies, one measure was theoretically superior to alternative measures because it made fuller use of available information. The analysis demonstrated that this measure had superior and excellent accuracy, with areas under the curve $>$ 0.9 for both weight and length.[36]

For some subfields, establishing the validity of measures is hampered by a lack of a definitive measure to use as a criterion to compare reference or field measures. Assessment of usual dietary intake, for example, presents many challenges to establishing validity.[39,40] In some situations, creative strategies can be used to develop definitive measures, as has been done for assessment of growth faltering,[36] child dietary intake,[13] and household or individual food insecurity (discussed below). When a definitive measure is not available for establishing the validity of a reference measure—or a reference measure is not available for establishing the validity of a field measure—then agreement with alternative measures can be established, but not validity.

Establishing cross-context equivalence requires examining the performance of a measure in multiple contexts to determine whether or not its performance is comparable. In particular, it is important to establish that its performance is consistent with the theory and it is precise, dependable, and accurate in different contexts. Establishing equivalence for subjective measures typically requires both cognitive interviewing and field testing.[2] For example, the measures used globally for assessing family care behaviors of young children in the Multiple Indicator Cluster Surveys were initially developed to be measurement equivalent on the basis of cognitive interviewing in seven countries and field testing in three countries[41]; further evidence demonstrating measurement equivalence has not been obtained. Cross-context equivalence often is assumed for objective measures such as weight and height on the assumption that the underlying physical, chemical, or biological process are universal, but the possibility that this assumption does not hold should be considered. For example, although body mass index obtained from weight and height measures on adults is associated with percentage of body fat in all global populations, the association differs by age, sex, and ethnic group.[42]

THREE EXAMPLES TO ILLUSTRATE CONCEPTS AND PRINCIPLES

Validation of Equation to Estimate Body Fat Mass from Skinfolds

The first example illustrates application of the concepts and principles for an objective measure using data from a study of 41 female crew, gymnastics, and track university athletes who were assessed for body fat mass using densitometry (ie, the reference measure) and skinfolds (ie, the field measures) by a well-trained observer (unpublished data). One purpose of the study was to establish the validity of using a combination of site-specific skinfold measures to estimate whole-body fat mass of individuals. Imprecision was estimated by taking the measures once and then repeating them several minutes later on the same individual. Undependability was estimated by taking two measures on the same individual 1 week apart, assuming that body composition did not change during that time. Both imprecision and undependability were calculated using standard deviations of differences, which is equivalent to fitting variance components models with the individual athlete as a random effect to separate out the components of among-athlete variance, undependability, and imprecision.

The imprecision and undependability of skinfold measurements at the biceps site were 0.22 and 0.09 mm, respectively; the thigh site had larger imprecision and undependability of 0.59 and 0.24 mm, respectively. These values for imprecision were at the low end of the typical range for imprecision at the biceps and thigh sites (0.2 to 0.6 mm and 0.5 to 0.7 mm, respectively).[43]

A multiple regression model for body fat mass on biceps and thigh skinfolds resulted in these estimates of the regression coefficients:

$$\text{Body fat mass} = 2.691 + (0.832 \times \text{biceps}) + (0.261 \times \text{thigh}) \tag{4}$$

with a standard deviation of the residuals of 2.21 kg and a multiple correlation of 0.864. In comparison, the standard deviation among individuals in the sample was 4.3 kg for body fat mass as measured by densitometry, so the standard deviation of the residuals that quantifies the error of prediction is large.

For estimating the mean body fat mass of a new sample of, say, 40 athletes all of average (relative to the original sample) biceps and thigh skinfolds, the 95% CI of the sample in the middle of the distribution using the standard formula for multiple linear regression was ±0.67 kg, which shows that the observed and predicted mean for a new sample of 40 would be expected to be fairly close to each other. The confidence interval would be wider for athletes with skinfolds smaller or larger than the middle of the distribution of the new sample. These results mean that this equation would provide an estimate of the average body fat mass in the new sample with reasonably useful combined accuracy and reliability, with a confidence interval spanning 1.34 kg.

The 95% CI for an estimated body fat mass of a new individual woman would be ±4.91 kg in the middle of the distribution. For the estimation of a new individual's body fat mass (ie, a different purpose), then, the combined accuracy and reliability was poor because the CI spanned nearly 10 kg. For comparison, the mean body fat mass of the women athletes in the study was 13.3 kg.

These results illustrate two points. First, judgment of validity depended on the purpose. Second, the correlation coefficient, which appeared high, was not helpful in determining whether the field measures had sufficient accuracy and reliability for the purposes envisioned. Furthermore, in the case that the underlying variation in body fatness in another sample (or a subsample of the same sample) of female athletes was higher (or lower), the correlation coefficient would be lower (or higher), even when the combined accuracy and reliability quantified in kilograms was the same, and therefore not useful in examining validity.

Validation of Household and Individual Food Insecurity

The second example illustrates application of the concepts and principles for a subjective measure using data that came from a study of 126 households followed over time in northern Burkina Faso.[44] This study used qualitative and quantitative methods to develop and validate an experience-based measure of household food insecurity for the purpose of assisting development organizations in evaluating the influence of programs intended to improve food security. The measure was developed assuming one underlying latent construct. Food insecurity in Burkina Faso has a strong seasonal pattern, and data were collected in five waves every July (hungry season) and January (postharvest season) from July 2001 to July 2003 to contrast the worst and best periods for food security. An experience-based food insecurity questionnaire was developed from the qualitative study following several steps.[2] To establish accuracy of this field measure, the measure from this questionnaire was compared to season (July vs January); a set of economic, dietary intake, and anthropometric measures; and a definitive measure developed for the study. For the definitive measure, the households were classified at two different times as to whether they were food secure, moderately food secure, or severely food insecure on the basis of the integrated, in-depth knowledge that a single observer had of each household's situation from multiple visits to each household, based on first principles about household food insecurity drawn from knowledge obtained in prior in-depth qualitative interviews in the province.

Validity was examined for three specific purposes: to capture overall seasonal differences, to discriminate among households at each wave, and to discriminate changes in households across waves. This example focuses on the second of these purposes. Each of the six criteria to establish validity for this purpose in the context of northern Burkina Faso was met.[44] Criteria 1 and 2 were met because the construction of the items for the questionnaire was based on the in-depth qualitative data, and the frequency of affirmative responses for the items was as expected based on the construction. Criteria 3 and 4 were met because the internal consistency at each wave was good (Cronbach α=.81 to .85), meaning that the measure was reliable. Criterion 5 was met because regression methods demonstrated that the household food insecurity measure was associated with the other measures usually indicative of household food insecurity in a pattern consistent with theory (eg, more strongly associated with dietary intake than with anthropometry). Furthermore, the household food insecurity measure was strongly associated with the definitive observer measure. Criterion 6, whether the performance of the household food insecurity score was attributable to its ability to capture household food insecurity status beyond that of other measures, was met as demonstrated by multiple multinomial logistic regression models using the observer classification as the definitive measure. Adding the household food insecurity measure to the model with the economic measures improved model fit significantly and increased the area under the receiver operating characteristic curve.

The method for developing the definitive measure was first demonstrated in a sample of households in New York State.[45] The method has been used successfully to develop definitive measures to help establish validity in several other studies of household[5,46,47] and child[37] food insecurity. These studies show the potential of creatively using qualitative methods to contribute to establishing validity of quantitative methods.

Experienced-based measures of household or individual food insecurity began to be developed in the late 1980s, and the validity of these measures for several purposes has been established from studies in the United States and many other countries.[3,8,48-51] Efforts to investigate equivalence of these measures across cultures and countries began about 10 years later. Coates and colleagues[52] examined whether there were cross-context commonalities of the food insecurity experience as captured in measures and ethnographies from 15 different countries. The study found that three core constructs (ie, insufficient food quantity, inadequate food quality, and uncertainty and worry about food) were experienced in all cultures, and that concerns about social unacceptability were present in all ethnographies. The relative frequency with which survey respondents affirmed items for the core constructs was similar across most cultures. Other research has also shown construct, item, and measurement equivalence of food-insecurity measures across countries.[3] Recently, the Food Insecurity Experience Scale was developed and fielded in 147 countries through the Gallup World Poll[53]; this scale was developed to be construct, item, and measurement equivalent, with subsequent analytic adjustment allowing comparisons of the scale scores and prevalence of categories across countries to be made.

Comparison of Two Field Measures of Fruit and Fruit Juice Consumption

The third example illustrates application of the concepts and principles for two subjective measures of fruit and fruit juice consumption using data from a study of 186 households in

Table. Frequency of fruit and fruit juice consumption from Behavioral Risk Factor Surveillance Survey (BRFSS) and 24-hour recall in a study of 186 women in rural Upstate New York in 1993[54]

			Variance		Standard Deviation		Reliability	
Measure	**Unit**	**Mean**	**Among individuals**	**Within individuals**	**1 d**	**Mean of 2 d**	**1 d**	**Mean of 2 d**
BRFSS	Square root of times per week	2.77[a]	2.02	—	1.42	—	—	—
24-h recall	Times per day	0.903[b]	0.493	1.08	1.25	1.02	0.311	0.477

[a] Mean in times per week estimated by square of 2.77=7.67.

[b] Mean of 2 days in times per week is 0.903×7=6.32.

rural Upstate New York.[54] First, two items from the Behavioral Risk Factor Surveillance Survey (BRFSS)[55] were used to ask women about usual frequency of consumption of fruits and fruit juices per week. Second, two quantitative 24-hour recalls were conducted about one month apart, and data from each recall was used to estimate the frequency of fruit and fruit juice consumption per day.

This example differs from the first two in that neither of the two measures (ie, fruit and fruit juice frequency from the BRFSS and from 24-hour recall) are definitive or reference measures for the purpose of estimating usual consumption of fruit and fruit juice of individuals in this rural population, and the study is of agreement rather than validity. Nevertheless, to facilitate comparison between the measures, the BRFSS measure is compared to the frequency of fruit and fruit juice consumption from 24-hour recall.

The test−retest reliability of the 1-day measure of frequency per day of the 24-hour recall of fruit and fruit juice frequency was 0.311 (Table). This reliability is the proportion of total variance due to the among-individual variance and is also the correlation of the values from the first and second recalls. Averaging measures for 2 days resulted in a 2-day reliability of 0.477 for frequency per day for the 24-hour recall, obtained by dividing the within-individual variance by 2 and recalculating the proportion of total variance due to the among-individual variance.

One purpose of the measures would be to estimate the mean frequency of fruit and fruit juice intake per week in the sample. In times per week, the means from the BRFSS and 24-hour recall were 7.67 and 6.32, respectively (Table). The lower mean for the 24-hour recall was expected given that, for estimating usual fruit and fruit juice consumption, the measure of frequency per day from the 24-hour recall was left-censored relative to the frequency per week from the BRFSS measure, reflecting episodic fruit and fruit juice consumption with clumping at zero (Figure 4). That is, not all zeros reported from 2 days of 24-hour recall were true zeros for usual intake over a week.

Two other purposes would be to differentiate either groups or individuals with high and low fruit and fruit juice intake. For these purposes, the combined accuracy and reliability of the BRFSS frequency per week for predicting the 24-hour recall frequency per day was examined using linear regression. Because the BRFSS data were skew, the square root of the

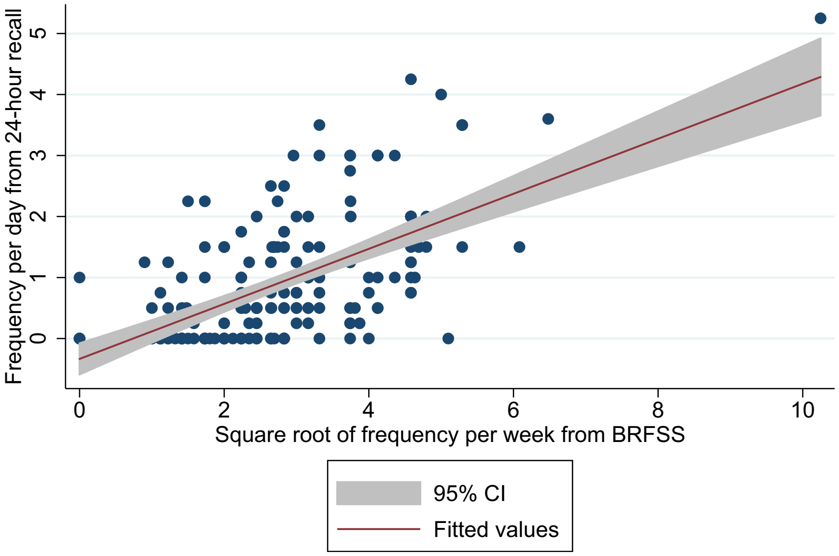

Figure 4. Frequency (times per day) of fruit and fruit juice consumption from 24-hour recall 2-day mean) vs square root of frequency (times per week) from Behavioral Risk Factor Surveillance Survey (BRFSS), with 95% CI for the mean, in study of 186 women in rural Upstate New York in 1993.[54]

frequency was used. Assuming a straight-line relationship was reasonable based on the scatterplot of the two measures (Figure 4). The regression model for the 24-hour recall frequency per day on the square root of BRFSS frequency per week resulted in these estimates of the regression coefficients:

$$\begin{aligned}&\text{24-hour recall frequency per day}\\&\quad = -0.334 + 0.451\sqrt{\text{BRFSS frequency per week}}\end{aligned} \tag{5}$$

with a standard deviation of the residuals of 0.805 times per day or 5.64 times per week, a correlation of 0.619, and a squared correlation of 0.383. To illustrate this equation, two individuals consuming fruit and fruit juice 1 time and 4 times per week (from BRFSS) differed on average by 0.451 times per day (from 24-hour recall) or 3.17 times per week. Thus, the BRFSS, compared with the 24-hour recall, appears to capture differences in absolute amounts similarly on average because 3.17 is close to 3. This closeness is evidence of agreement for differences on average. Both measures may be prone to bias.

The 95% CI for the prediction of the mean of a hypothetical new sample of 186 from the same location as the original sample was about ±0.116 times per day (see Figure 4) or ±0.812 times per week. The 95% CI for the prediction of a new single individual was ±1.58 times per day or ±11.1 times per week. As with the example of body fat mass in the first example and as reflected in the considerable scatter of the points around the regression line in Figure 4, the prediction of an individual's frequency of fruit and fruit juice consumption per day from the BRFSS frequency measure would be much less accurate as compared to the 24-hour recall than the prediction of the average of the group.

The widths of these confidence intervals were inflated because the reliability of 2 days of recall was slightly less than half. Based on these estimates, hypothetically in the case that a large number of days of recall (ie, a few hundred) were taken to remove the day-to-day undependability and imprecision (ie, unreliability), the standard deviation of the residuals would be 0.310 times per day, the CI would be ±0.609 times per day, and the correlation would be 0.892.

The residual plot shows reasonable homoscedasticity (ie, equal variability) except at the left side where the censoring occurred, with the 24-hour recall measure of frequency per day being mildly skew (Figure 5). The residual plot is similar to a Bland-Altman plot[33]; the former plots the difference against the predictor measure whereas the latter plots the difference against the average of the two measures.

DISCUSSION

Lack of clarity about terminology used when discussing validity of measures has been a persistent problem in the literature. This article explains the concepts and terminology of the biometric conceptual system for validity that is well established in the chemical and biological sciences. The biometric conceptual system is extensively used in the field of nutritional sciences to establish validity of anthropometry, other objective measures, and some subjective measures (eg, household and individual food insecurity).[8] This system is well grounded, coherent, and complete conceptually, and articulated with statistical models for quantifying components of error. An alternative psychometric conceptual system is commonly used among social and behavioral scientists,[16,18,19] including in nutritional sciences, and the alignment of the two systems was shown.

Although inaccuracy and unreliability are both important components of error, scientists tend to put more weight on accuracy than reliability. One reason is that sometimes it is easier to improve reliability in usage (eg, by taking the average of repeated measures) than it is to improve accuracy. Nevertheless, the simple decomposition of total measurement error into components of inaccuracy and unreliability

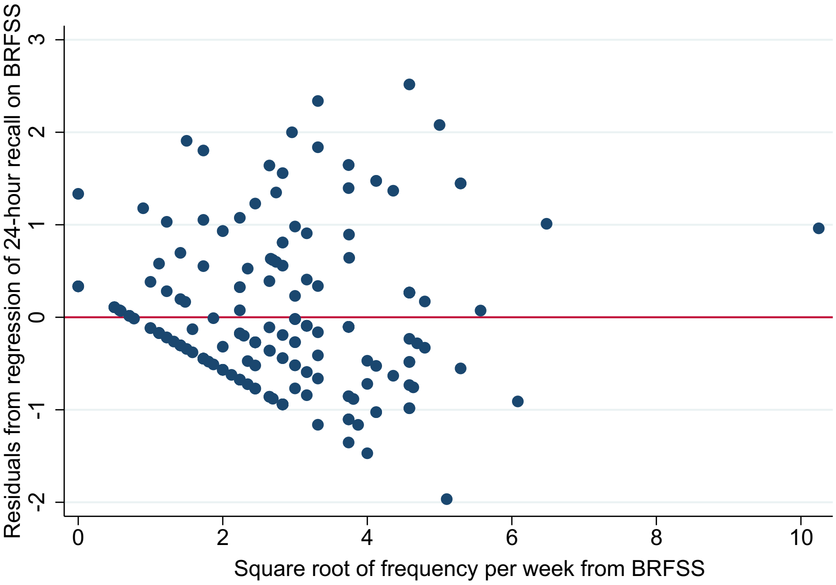

Figure 5. Residuals from regression of frequency (times per day) of fruit and fruit juice consumption from 24-hour recall (2-day mean) on square root of frequency (times per week) from Behavioral Risk Factor Surveillance Survey (BRFSS), in study of 186 women in rural Upstate New York in 1993.[54]

reminds us that the aim should be to reduce total error, and that sometimes that may occur by choosing a measure that is highly reliable although somewhat inaccurate.

Being able to make these decisions requires estimation of both inaccuracy and unreliability, and then judgment about whether these errors are sufficiently small to provide useful analytic measurement for a particular purpose and context. Simple correlation coefficients are inadequate in most instances as a basis for such judgment, and estimates of absolute errors are superior to relative errors.

A substantial limitation of the literature is the pervasiveness of claims that a measure is valid—as if validity is inherent to that measure. Rather, validity refers to the suitability of a measure when used for a particular purpose in a particular context, and so specifying the purpose and context is necessary when making a claim for validity. A measure shown to be valid for a specific purpose and context may not be valid for other purposes or contexts; that is, validity must be demonstrated for each intended particular purpose. An important implication is that statements such as a "measure is validated" or that "we used a validated measure" are not appropriate. Future reports that refer to validity should instead make a statement such as "measure A has been shown to be valid for assessing construct B for purpose C in context D (ie, circumstances of the measurement situation such as population and culture) by comparing it with (reference or definitive) measure E, resulting in precision F, dependability G, and accuracy H (ideally all specified in absolute units)." For future reports that refer to agreement, the statement becomes: "measure J has been shown to achieve agreement K when compared to measure L for assessing construct M for purpose N in context O."

In the case that a measure is intended to be used across multiple contexts for comparative purposes, specific efforts to establish cross-context equivalence as well as validity are needed. Determining how much evidence of validity and cross-context equivalence is sufficient to justify using a measure is difficult to state succinctly. The emphasis in the definition of validity on whether a measure is useful is important in this regard. Judgment of usefulness involves technical as well as other scientific and policy appraisal. For example, the United Nations Children's Fund implemented in the fourth round of the Multiple Indicator Cluster Surveys a set of 10 questionnaire items to assess, through parental report, the early childhood development of 36- to 59-month-old children. Although the development and testing of the items was done through a series of steps, the development and validation process was limited because of the compelling need to place some items in the survey given the policy importance of early childhood development and the lack of any epidemiologic data. Although some of the items (and resulting indicators) that were implemented have weaknesses, the set of items has allowed for the first time global estimates of prevalence of inadequate early childhood development[56] and epidemiologic demonstration of the importance of family care behaviors to early childhood development.[57] Further efforts have been undergoing processes to develop, test, and implement improved items for future surveys, but the first effort demonstrated the usefulness of measures and indicators from survey items for addressing policy questions.

A related practical decision faced often by scientists is whether to accept existing evidence of validity and equivalence as opposed to needing to further establish these. This decision is particularly challenging when using subjective measures that depend on communication between interviewer and respondents and reported by respondents. For subjective measures, ideally there will be evidence obtained in multiple contexts from in-depth qualitative research about the important constructs and how to convey them, cognitive interviewing about whether candidate items are understood by respondents as intended, and field testing to provide quantitative data to examine reliability and accuracy.[2] In the case that there is sufficient evidence to establish validity for a given purpose in multiple contexts or at least one that is similar to the one being considered—and there is reason to expect that the underlying property being measured and the measuring processes are shared across contexts in a common way—then it may be justified to assume that a version of the measure that is adapted to the current context will be suitable to provide useful measurement.

When planning a study to provide evidence about the validity of a measure, the first step should always be to clarify the purpose and context for use of that measure for which evidence of validity is sought because the requirements to establish validity will differ with the purpose. Any of several purpose(s) may be intended at the group or individual level or both (Figure 2). Second, one of the two conceptual systems for validation (Figure 3) should be chosen, and the concepts and terminology of that system adopted. Third, regardless of the system chosen, the validation study should seek to establish the extent to which the measure is well constructed (ie, grounded in an understanding of the underlying phenomenon of interest), reliable (ie, produces precise and dependable data), and accurate (ie, produces data representative of the true value, free from bias). Establishing accuracy requires that at least one other measure understood to be more accurate be available for comparison (ie, a reference or definitive measure when validating a field measure or a definitive measure when validating a reference measure). In the case that no other measure understood to be more accurate is available, then the study will be able to establish agreement rather than validity. Fourth, analytic methods that go beyond simple correlations are needed to provide a basis for making reasoned judgment about whether the measure provides useful analytic measurement for the particular purpose(s) and context. A technical guide[2] is available that provides detailed steps for the development and validation of a measure that could be helpful in planning validation studies.

References

1. Wernimont G. Statistical Control of Measurement Processes. In: Gould RF, ed. *Validation of the Measurement Process*. Washington, DC: American Chemical Society; 1977:1-29.
2. Frongillo EA, Nanama S, Wolfe WS. *Technical Guide to Developing a Direct, Experience-Based Measurement Tool for Household Food Insecurity*. Washington, DC: Food and Nutrition Technical Assistance, Academy for Educational Development; 2004.

3. Leroy JL, Ruel M, Frongillo EA, Harris J, Ballard TJ. Measuring the food access dimension of food security: A critical review and mapping of indicators. *Food Nutr Bull*. 2015;36(2):167-195.
4. Riely F, Mock N, Cogill B, Bailey L, Kenefick E. *Food Security Indicators and Framework for Use in the Monitoring and Evaluation of Food Aid Programs*. Arlington, VA: Food Security and Nutrition Monitoring Project (IMPACT), ISTI, Inc, for the US Agency for International Development; 1999.
5. Wolfe WS, Frongillo EA, Valois P. Understanding the experience of food insecurity by elders suggests ways to improve its measurement. *J Nutr*. 2003;133(9):2762-2769.
6. Koch DD. Concepts in the validation of neurochemical methods: The proper generation and use of statistics. *Life Sci*. 1987;41(7):853-856.
7. Habicht J-P, Yarbrough C, Martorell R. Anthropometric field methods: Criteria for selection. In: Jelliffe DB, Jelliffe EFP, eds. *Nutrition and Growth (A Comprehensive Treatise)*. Vol 2. Boston, MA: Springer; 1979:365-387.
8. Frongillo EA. Validation of measures of food insecurity and hunger. *J Nutr*. 1999;129(2 suppl):506S-509S.
9. Lampl M, Birch L, Picciano MF, Johnson ML, Frongillo EA. Child factor in measurement dependability. *Am J Hum Biol*. 2001;13(4):548-557.
10. Bell RC, Lanou AJ, Frongillo EA, Levitsky DA, Campbell TC. Accuracy and reliability of total body electrical conductivity (TOBEC) for determining body composition of rats in experimental studies. *Physiol Behav*. 1994;56(4):767-773.
11. Uriano GA, Cali JP. Role of reference materials and reference methods in the measurement process. In: Gould RF, ed. *Validation of the Measurement Process*. Washington, DC: American Chemical Society; 1977:140-161.
12. Ellis KJ. Human body composition: In vivo methods. *Physiol Rev*. 2000;80(2):649-680.
13. Diep CS, Hingle M, Chen TA, et al. The automated self-administered 24-hour dietary recall for children, 2012 version, for youth aged 9 to 11 years: A validation study. *J Acad Nutr Diet*. 2015;115(10):1591-1598.
14. Freedman LS, Commins JM, Moler JE, et al. Pooled results from 5 validation studies of dietary self-report instruments using recovery biomarkers for energy and protein intake. *Am J Epidemiol*. 2014;180(2):172-188.
15. Oreskes N, Shrader-Frechette K, Belitz K. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*. 1994;263(5147):641-646.
16. Gleason PM, Harris J, Sheean PM, Boushey CJ, Bruemmer B. Publishing nutrition research: Validity, reliability, and diagnostic test assessment in nutrition-related research. *J Am Diet Assoc*. 2010;110(3):409-419.
17. Kipnis V, Subar AF, Midthune D, et al. Structure of dietary measurement error: Results of the OPEN biomarker study. *Am J Epidemiol*. 2003;158(1):14-21.
18. Nunnally JC. *Psychometric Theory*. 2nd ed. New York, NY: McGraw-Hill; 1978.
19. Kelly P, Fitzsimons C, Baker G. Should we reframe how we think about physical activity and sedentary behaviour measurement? Validity and reliability reconsidered. *Int J Behav Nutr Phys Act*. 2016;13(1):32.
20. Churchill GA. A paradigm for developing better measures of marketing constructs. *J Mark Res*. 1979;16:64-73.
21. Byrne BM, Watkins D. The issue of measurement invariance revisited. *J Cross Cult Psychol*. 2003;34(2):155-175.
22. Vijver FV. Bias and equivalence: Crosscultural perspectives. In: Harkness J, Vijver FV, Mohler P, eds. *Cross-Cultural Survey Methods*. Hoboken, NJ: Wiley-Interscience; 2003:143-155.
23. Coates J, Swindale A, Bilinsky P. Household Food Insecurity Access Scale (HFIAS) for Measurement of Household Food Access: Indicator Guide. Version 3. Washington, DC: Food and Nutrition Technical Assistance Project, Academy for Educational Development; 2007.
24. Deitchler M, Ballard T, Swindale A, Coates J. *Validation of a Measure of Household Hunger for Cross-Cultural Use*. Washington, DC: Food and Nutrition Technical Assistance II Project; 2010.
25. Gregorich SE. Do self-report instruments allow meaningful comparisons across diverse population groups?: Testing measurement invariance using the confirmatory factor analysis framework. *Med Care*. 2006;44(suppl):S78-S94.
26. Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organ Res Methods*. 2000;3(1):4-70.
27. Schmitt N, Kuljanin G. Measurement invariance: Review of practice and implications. *Hum Resour Manage Rev*. 2008;18:210-222.
28. Fuller WA. *Measurement Error Models*. New York, NY: John Wiley & Sons; 1987.
29. Groves RM. Measurement error across the disciplines. In: Biemer PP, Groves RM, Lyberg LE, Mathiowetz NA, Sudman S, eds. *Measurement Error in Surveys*. Hoboken, NJ: John Wiley & Sons; 1991.
30. Biemer PP, Stokes SL. Approaches to the modeling of measurement errors. In: Biemer PP, Groves RM, Lyberg LE, Mathiowetz NA, Sudman S, eds. *Measurement Errors in Surveys*. Hoboken, NJ: John Wiley & Sons; 1991.
31. Cappelleri JC, Lundy JJ, Hays RD. Overview of Classical Test Theory and Item Response Theory for quantitative assessment of items in developing patient-reported outcome measures. *Clin Ther*. 2014;36:648-662.
32. Snedecor GW, Cochran WG. *Statistical Methods*. 7th ed. Ames, IA: The Iowa State University Press; 1980.
33. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307-310.
34. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med*. 1990;20(5):337-340.
35. Hebert JR, Miller DR. The inappropriateness of conventional use of the correlation coefficient in assessing validity and reliability of dietary assessment methods. *Eur J Epidemiol*. 1991;7(4):339-343.
36. Frongillo EA, Rothe GE, Lambert JKJ. Determining growth faltering with a tracking score. *Am J Hum Biol*. 1990;2(5):491-501.
37. Fram MS, Frongillo EA, Draper CL, Fishbein EM. Development and validation of a child report assessment of child food insecurity and comparison to parent report assessment. *J Hunger Environ Nutr*. 2013;8(2):128-145.
38. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988;240(4857):1285-1293.
39. Hébert JR, Hurley TG, Steck SE, et al. Considering the value of dietary assessment data in informing nutrition-related health policy. *Adv Nutr*. 2014;5(4):447-455.
40. Subar AF, Freedman LS, Tooze JA, et al. Addressing current criticism regarding the value of self-report dietary data. *J Nutr*. 2015;145(12):2639-2645.
41. Kariger P, Frongillo EA, Engle P, Britto PMR, Sywulka SM, Menon P. Indicators of family care for development for use in multicountry surveys. *J Health Popul Nutr*. 2012;30(4):472-486.
42. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet*. 2004;363(9403):157-163.
43. Harrison GG, Buskirk ER, Lindsay Carter JE, et al. Skinfold thicknesses and measurement technique. In: Lohman TG, Roche AF, Martorell R, eds. *Anthropometric Standardization Reference Manual*. Champaign, IL: Human Kinetics Books; 1988.
44. Frongillo EA, Nanama S. Development and validation of an experience-based measure of household food insecurity within and across seasons in Northern Burkina Faso. *J Nutr*. 2006;136(5 suppl):1409S-1419S.
45. Frongillo EA, Rauschenbach BS, Olson CM, Kendall A, Colmenares AG. Questionnaire-based measures are valid for the identification of rural households with hunger and food insecurity. *J Nutr*. 1997;127(5):699-705.
46. Wolfe WS, Olson CM, Kendall A, Frongillo EA. Hunger and food insecurity in the elderly: Its nature and measurement. *J Aging Health*. 1998;10(3):327-350.
47. Hamelin A-M, Beaudry M, Habicht J-P. Characterization of household food insecurity in Québec: Food and feelings. *Soc Sci Med*. 2002;54(1):119-132.

48. Studdert LJ, Frongillo EA, Valois P. Household food insecurity was prevalent in Java during Indonesia's economic crisis. *J Nutr*. 2001;131(10):2685-2691.
49. Frongillo EA, Chowdhury N, Ekström E-C, Naved RT. Understanding the experience of household food insecurity in rural Bangladesh leads to a measure different from that used in other countries. *J Nutr*. 2003;133(12):4158-4162.
50. Coates J, Wilde PE, Webb P, Rogers BL, Houser RF. Comparison of a qualitative and a quantitative approach to developing a household food insecurity scale for Bangladesh. *J Nutr*. 2006;136(5 suppl):1420S-1430S.
51. Melgar-Quinonez HR, Zubieta AC, MkNelly B, Nteziyaremye A, Gerardo MFD, Dunford C. Household food insecurity and food expenditure in Bolivia, Burkina Faso, and the Philippines. *J Nutr*. 2006;136(5 suppl):1431S-1437S.
52. Coates J, Frongillo EA, Rogers BL, Webb P, Wilde PE, Houser R. Commonalities in the experience of household food insecurity across cultures: What are measures missing? *J Nutr*. 2006;136(5 suppl):1438S-1448S.
53. Cafiero C, Nord M, Viviani S, et al. *Voices of the Hungry: Methods for Estimating Comparable Prevalence Rates of Food Insecurity Experienced by Adults in 147 Countries*. Rome, Italy: Food and Agriculture Organization; 2015.
54. Kendall A, Olson CM, Frongillo EA. Relationship of hunger and food insecurity to food availability and consumption. *J Am Diet Assoc*. 1996;96(10):1019-1024.
55. Wolfe WS, Frongillo EA, Cassano PA. Evaluating brief measures of fruit and vegetable frequency and variety: Cognition, interpretation, and other measurement issues. *J Am Diet Assoc*. 2001;101(3):311-318.
56. McCoy DC, Peet ED, Ezzati M, et al. Early childhood developmental status in low- and middle-income countries: National, regional, and global prevalence estimates using predictive modeling. *PLOS MED*. 2016;13(6):e1002034.
57. Frongillo EA, Kulkarni S, Basnet S, de Castro F. Family care behaviors and early childhood development in low- and middle-income countries. *J Child Fam Stud*. 2017;26(11):3036-3044.

For more information on the subject discussed in this article, see Sites in Review on page 1967.

AUTHOR INFORMATION

E. A. Frongillo is a professor, Department of Health Promotion, Education, and Behavior, Arnold School of Public Health, University of South Carolina, Columbia. T. Baranowski is a professor, Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX. A. F. Subar is acting branch chief, Risk Factor Assessment Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD. J. A. Tooze is a professor, Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC. S. I. Kirkpatrick is an assistant professor, School of Public Health and Health Systems, University of Waterloo, Waterloo, Ontario, Canada.

Address correspondence to: Edward A. Frongillo, PhD, Department of Health Promotion, Education, and Behavior, 915 Greene St, Discovery I, University of South Carolina, Columbia, SC 29208-4005. E-mail: efrongil@mailbox.sc.edu

STATEMENT OF POTENTIAL CONFLICT OF INTEREST

No potential conflict of interest was reported by the authors.

FUNDING/SUPPORT

This research was partially funded by institutional support from the US Department of Agriculture, Agricultural Research Service (Cooperative Agreement 58-3092-5-001).

AUTHOR CONTRIBUTIONS

E. A. Frongillo wrote the first draft. All authors reviewed and commented on subsequent drafts of the manuscript.