Citation in APA format: Hoyle, R. H., & Borsboom, D., & Tay, L. (2025). Measuring constructs. In D. T. Gilbert, S. T. Fiske, E. J. Finkel, & W. B. Mendes (Eds.), *The handbook of social psychology* (6th ed.). Situational Press. https://doi.org/10.70400/OUQF7656

Measuring Constructs

Rick H. Hoyle, Duke University

Denny Borsboom, University of Amsterdam

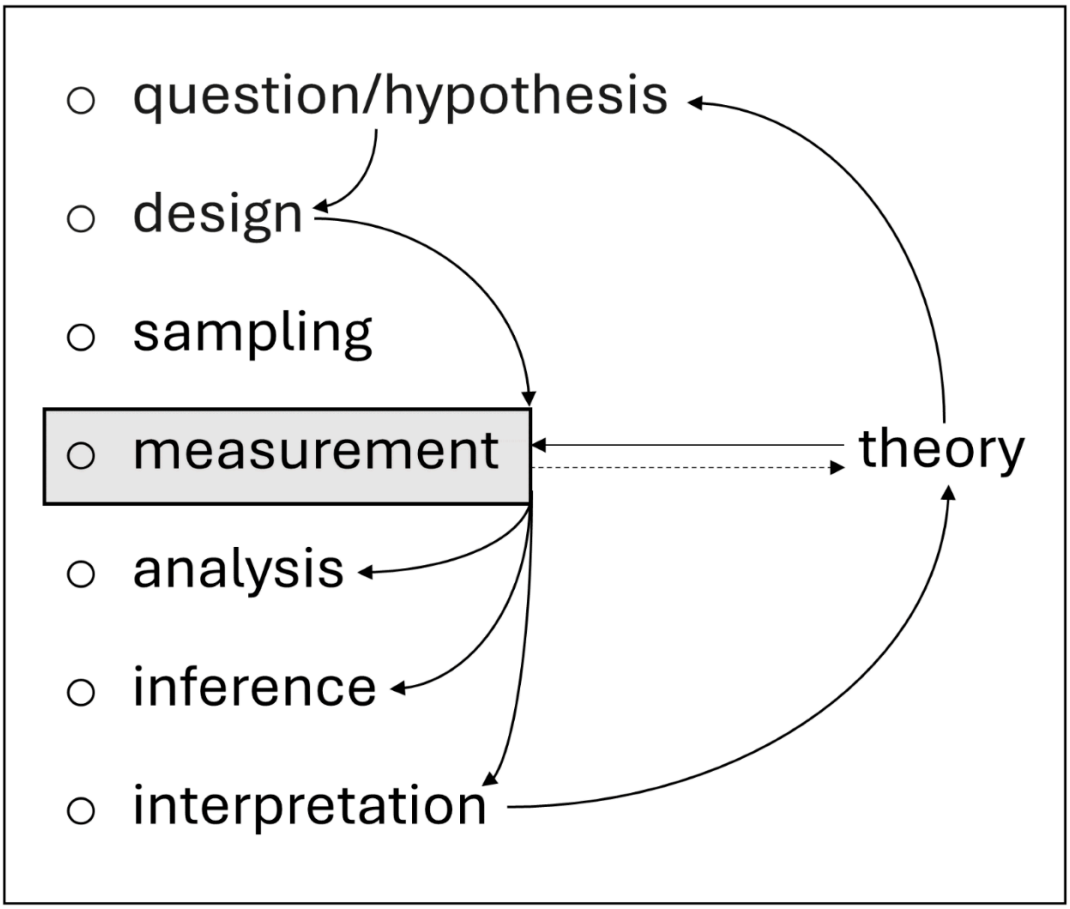
Louis Tay, Purdue University

> "Science cannot progress without reliable and accurate measurement of what it is you are trying to study. The key is measurement, simple as that." --Robert D. Hare[1]

Reliable and accurate measurement of well-delineated constructs is essential for rigorous and replicable research in social psychology, regardless of the research question, research design, method of data collection, or data analytic strategy. Measurement in social psychology can take many forms, all of which aim to differentiate people, objects, or situations according to their standing on attributes that define constructs posited by theories. That differentiation is characterized by two or more labels or numeric values corresponding to different quantities, intensities, expressions, or forms of the construct. The manipulation and analysis of these products of measurement allow for informative descriptions of social-psychological phenomena and tests of hypotheses and models that implicate the constructs they represent. Without valid measurement, these essential research activities are not possible, and the scientific goals that motivate them are not achievable.

Despite the fundamental relevance of measurement for all research in social psychology, it is poorly understood and often poorly done (Flake & Fried, 2020). Rather, multiple critiques of research in social psychology and recommendations for improving it have focused on rigorous research design, careful sampling, and properly applied and interpreted statistical methods (e.g., Bakker & Wicherts, 2011; Borgatta & Bohrnstedt, 1974; Brewer & Crano, 2014; Cialdini, 2009; Henry, 2008). Although important, excellence in all of these research components cannot overcome the consequences of poor measurement; better measurement is one of the most promising paths to more informative, replicable, and generalizable research results in social psychology. As illustrated in Figure 1, measurement is a critical nexus in the conduct of empirical research in social psychology. It is informed by theory (and may directly inform it; Greenwald, 2012), the motivating question or hypothesis, and research design. It informs analysis, inference, and interpretation, which inform theory. Given its centrality in the research process, it is surprising there have been few critiques of measurement relevant to social psychology and recommendations for addressing them (Fishman et al., 2021; Flake & Fried, 2020; Kornbrot et al., 2018; Lilienfeld & Strother, 2020). The fundamental relevance of measurement for all research by social psychologists requires that it be a central focus of any critique of the field or attempts to improve it.

Figure 1: The Central Role of Measurement in Social-Psychological Research



To that end, the overarching goal of this chapter is to promote a deeper understanding of measurement in social psychology and an appreciation for the challenge and necessity of doing it well. The core content of the chapter begins with a brief discussion of measurement theory, highlighting the specific challenges associated with measuring constructs. Then, to emphasize the necessity of well-delineated constructs for effective measurement, the next major section discusses construct theories, emphasizing their implications for measurement. Building on this foundation, the focus moves to practical and strategic considerations for designing, using, and interpreting results from measures of constructs typical of research in social psychology. The presentation emphasizes the diversity of measurement approaches and key considerations when using them. Because of its widespread use in social psychology, the chapter gives special attention to self-report measurement. The final section highlights the properties of scores generated by measurement and their evaluation. A deeper understanding of measurement in social psychology and the theoretical foundation on which it rests will result in greater attention to and care in measuring constructs.

THEORIES AND MODELS OF MEASUREMENT

The term *measurement* is laden with controversy. The psychometrician Louis Guttman confessed, in his presidential address to the Psychometric Society, that he had "avoided the term 'measurement' in all . . . writing and teaching" as it had "proved to arouse many irrelevant associations for various people" that form "actual barriers to progress in theory construction and in research" (Guttman, 1971, p. 330). The question of what measurement is and whether it can be achieved still stimulates controversy and debate. Whereas some scholars argue that measurement is just a matter of putting

numbers on a set of observations in a structured way (Stevens, 1946), others believe these observations should adhere to particular axioms (Luce et al., 1990). Some argue that measurement must involve estimating quantities (Michell, 1997, 1999), whereas others include determining ordered or unordered categories under the measurement umbrella (Sijtsma & Van der Ark, 2020). Some continue to hope for a breakthrough in measurement practice that would realize "the revolution that never happened" (Cliff, 1992) through the identification of axiomatic and model-fitting strategies that make psychological measurement more lawful and scores that result from measurement more interpretable and meaningful. Others claim that such a revolution can never happen because the subject matter of psychology, social psychology in particular, focuses on variables that are not amenable to direct manipulation, making lawful and meaningful measurement de facto impossible (Trendler, 2009; but see Markus & Borsboom, 2012).

The fact that these debates continue in the methodological literature may be surprising to many social psychologists, who purport to measure attitudes, emotions, and behavior regularly. Although intervals and ratios are rare in practice, nominal and ordinal scales exist in abundance. Measurement issues are addressed as matters of validity and reliability. Whether these practices are sufficient for realizing meaningful measurement cannot easily be determined. Social psychologists routinely collect responses from people and convert those responses to numerical values, using direct functions such as a total score on an attitude questionnaire or advanced statistical strategies (e.g., factor analysis). Data are collected to assess what level of precision can be achieved (Mellenbergh, 1996) and whether the scores obtained can be interpreted as desired (Kane, 2006). The question under debate, however, is whether these practices meet the methodological criteria necessary to justify the claim that their use actually realizes measurement (Briggs, 2022; Michell, 1997).

A further complication is that methodological criteria involve epistemic (and sometimes non-epistemic) values (Wijsen et al., 2022) and are, therefore, to some extent up for debate. Measurement is a particularly salient topic of discussion among methodologists because it is the centerpiece of science—the precise location in the research process at which theories and observations make contact. And because it is such a centerpiece, the semantics of measurement differ depending on assumptions about what constitutes empirical science and the overarching philosophy of science on which those assumptions rest (Borsboom, 2005). For a realist, who thinks that the goal of science is to provide true explanatory theories by "carving nature at its joints" (Meehl, 1978; Popper, 1959), measurement is a method to approximate real psychological attributes (Borsboom et al., 2004). For an empiricist, who thinks that the goal of science is to provide parsimonious yet accurate descriptions of observations (van Fraassen, 1980), measurement is a method to represent the outcome of various assessment methods and experimental manipulations (Luce et al., 1990). For a pragmatist, who believes that science aims to provide useful tools to predict and control consequential outcomes, measurements are as good as the predictions and manipulations they allow (Torres Irribarra, 2021). Thus, the semantics of measurement are not settled.

COMPONENTS OF MEASUREMENT

Acknowledging the tension between these competing views, this section highlights important themes from multiple theories of and approaches to psychological measurement. First, measurement aims to represent differences in characteristics of interest in meaningful and useful ways. Second, measurement is typically subject to random and systematic error that must be addressed, often through statistical modeling. Third, measurement forges a connection between

constructs and scientific observations, and therefore, it must have a substantive theoretical component.

Representational Component

Measurement invariably involves assigning symbols (numerals, letters, labels) according to some empirical procedure (Stevens, 1946). For instance, suppose a researcher studying attitudes towards abortion counts the number of pro-abortion statements with which respondents agreed in a survey. When scores are encoded into numbers, they represent the raw observations using the concept of order: People with higher scores endorsed more pro-abortion statements. However, in another empirical situation, a researcher may choose categorical representations; an example is psychopathology research, in which people are categorized as "cases" or "controls." In this instance, the assigned symbols represent a qualitative feature of the observations: whether they match a set of diagnostic criteria. Measurement procedures always contain such a representational step.

During the 20th century, the representational features of measurement were elevated into a general theory of measurement. *Representational Measurement Theory* (RMT) posits that measurements are essentially representations of possible and actual empirical relations between objects (Michell, 1993). In this view, when one states "object x measures 2 centimeters," one is not assigning the object a property; instead, one is communicating many empirical relations between the object and other objects in mathematical shorthand. For example, "if one laid two objects of 1 cm length head to head, the resulting composite y would not be noticeably different in length from object x." Because there are infinitely many combinations that produce an object of 2 centimeters, and infinitely many in which such an object can take part, the sentence "object x is 2 centimeters long" expresses the truth of an infinity of empirical statements, as does any other statement involving the length of an object. So how does this work? What should this infinity of relations between objects be like for them to be representable by simply assigning a single number to an object? What are the necessary and sufficient conditions for such a representation to succeed, and are those conditions met in psychological measurement?

RMT, as it evolved through the work of scholars like Campbell (1920), Nagel and Hempel (1931), Stevens (1946), and Suppes and Zinnes (1963), eventually culminated in an answer to this question in *Foundations of Measurement* (Krantz et al., 1971, 1989; Suppes et al., 1989). *Foundations* covers an impressive variety of measurement systems, for which the authors ask two questions. First, what conditions among empirical relations should hold between objects, so that a particular way of assigning numbers to objects can represent them? Second, how can these assigned numbers be transformed and maintain this representation? The first question is answered through a *representation theorem*, a mathematical proof that gives conditions under which the numerical assignment can represent empirical relations present in a system. The second question is answered by a *uniqueness theorem*, a mathematical proof that identifies the class of transformations that leave the representation intact.

If a representation and uniqueness theorem can be given for a numerical assignment, then one knows the level of measurement achieved (Stevens, 1946). If only multiplication by a positive constant is allowed, then the scale is a ratio scale; if general linear transformations like $x' = a + bx$ are also allowed, then one has an interval scale; if all monotonic transformations are also allowed, then one has an ordinal scale; and if all one-to-one transformations are allowed, then one has a nominal scale. This information is necessary, but not sufficient, as one cannot assess scale levels if no

representation theorem has been supplied. Because the great majority of measurement situations in social psychology have no such theorem, which scale level measurements instantiate (if any) cannot easily be determined (Markus & Borsboom, 2013).

There are several reasons why RMT has received little attention in psychological measurement (see also Cliff, 1989) despite the rigorous and lawful approach to measurement it posits. First, RMT, as it is typically presented, requires an assessment of the relations between objects—the empirical relations must be determined independently of the numerical assignment. However, in social psychology, researchers typically cannot determine whether two people hold the same attitude, have the same level of an emotion, or are equally moral independently of numerical assignments. Moreover, procedures are not available to manipulate them experimentally to the required level of precision (Trendler, 2009). Second, in many cases, it is difficult to construct a measurement system without invoking, a priori, the hypothesis that one is working with the right kind of attribute (e.g., a continuous linear order). In many cases, measurement outcomes have to be assessed, interpreted, and corrected using theories that require an account of how the measurement instrument works; such a theory itself typically invokes a hypothesis of how the targeted attribute is structured (Tal, 2021). Third, observations of human beings are noisy—they can be inconsistent over time and situations in ways that are difficult to understand. As a result, measurement models in psychology are assumed to require a statistical, probabilistic component to represent these error structures. Finally, it is not trivial to assess whether the axioms of RMT are satisfied (although some argue that a particular psychometric model known as the Rasch model embodies them; Perline et al., 1979, but also see Borsboom & Zand Scholten, 2008, and Kyngdon, 2008). Nevertheless, a basic understanding of representational measurement makes clear why additional components of measurement in social psychology are necessary. It also highlights the advantages of scores produced by measurement that have well-understood properties that make them useful, interpretable, and actionable—they have numerical traceability (Uher, 2022)

Statistical Component

Empirical researchers generally share the view that the representations that result from measurement (a total score, or a categorization) may be tainted by a variety of factors that are not of substantive interest. These factors are often collectively classified by terms such as “noise” or “measurement error” (van Bork et al., 2024; Viswanathan, 2005). Following this reasoning, there is a difference between the observed scores and the scores if the noise were removed. However, even if the noise cannot be removed altogether, it can be addressed by constructing a statistical model that places assumptions that encode the expected properties of the noise. For example, assume that noise comprises the sum of many independent factors specific to the particular circumstance in which a questionnaire item was encountered (e.g., noise or other distractions, the presence of other people, confusing or incomplete instructions). In that case, various statistical models that operationalize the concept of “random error” using a normal distribution can be used, because any sum of independent factors will approach a normal distribution as the number of factors grows (Fischer, 2011). A theory that is entirely based on such assumptions about error is *classical test theory* (Lord & Novick, 1968), which is used to conceptualize an important theoretical property known as the true score (defined as the expectation of the observed score). It is the basis for estimates of reliability commonly used by social psychologists, defined as the ratio of variance in true scores to variance in observed scores on a measure. The typically observed values less than 1.0 would, under the assumptions of classical test theory, indicate the presence of measurement error (i.e., unreliability).

Substantive Theoretical Component

A critical concern of measurement is how scores connect to the constructs they were intended to represent. In most approaches in psychology, this relates to how to connect the representational and statistical components. For instance, even if one counts the number of endorsed pro-abortion items, one could still hypothesize the actual attitude to be categorical (e.g., pro or con). And even if one matches scores to diagnostic criteria, one might still believe that the actual underlying construct of depression is dimensional (Kotov et al., 2021). As such, the relation between the theoretical construct and the observations becomes critically important (Borsboom et al., 2003, 2004; Cronbach & Meehl, 1955). This question cannot be answered by purely methodological means; rather, it requires one to actively theorize about what terms like "attitude" designate and how their referents are causally relevant to the measurement outcomes (De Houwer et al., 2009).

Thus, taking stock, researchers (a) aim to represent structures in data produced by measurement, but because (b) the data themselves are contaminated by noise, (c) researchers, lacking a representational component, require a substantive theory to connect the data structures to the targeted construct. In principle, there are many ways in which such a relation can be forged, and these define different approaches to measurement (Markus & Borsboom, 2013).

CONSTRUCTS

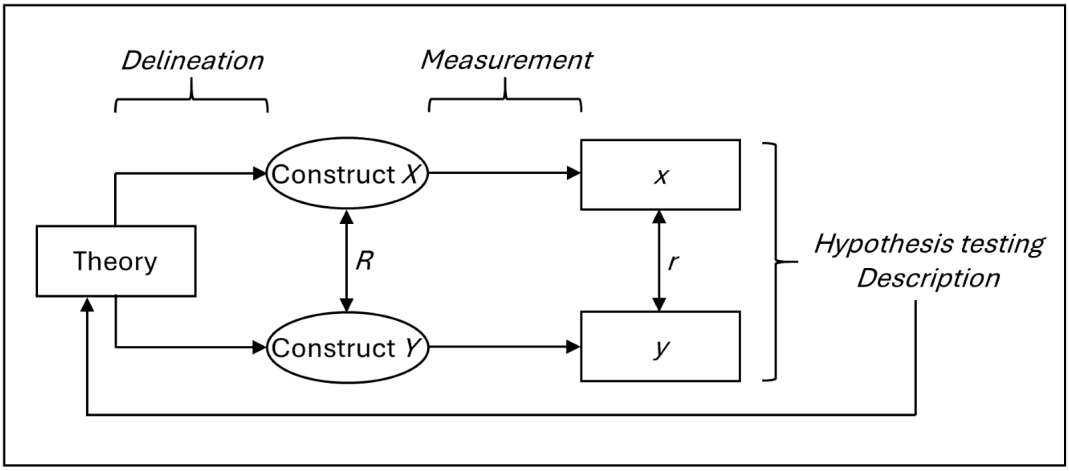
The purpose of measurement in social psychology is to represent constructs in ways that allow for description, modeling, and hypothesis testing in the service of explanation and prediction. For this reason, effective development, evaluation, and use of measures always require characterizing the properties of the construct they aim to represent (Cronbach & Meehl, 1955). Explicit and clear delineation of constructs is therefore essential for effective measurement in social psychology. That delineation and its measurement reflect rarely considered assumptions about the nature of constructs commonly referenced in theoretical accounts of social phenomena (for an informative history and review of constructs in psychology, see Slaney & Racine, 2013).

An Expanded Definition

In their seminal paper on construct validity, Cronbach and Meehl (1955) proposed that, "A construct is some postulated attribute of people, assumed to be reflected in test performance" (p. 283; "test performance" here refers to responses to or performance on measures of psychological attributes). To encompass the full range of targets, attributes, and forms of constructs and sources of information about them covered below, a broader definition is necessary, one that extends beyond people and tests. Moreover, the constructs typical of social-psychological theory and measurement frequently include multiple attributes. Thus, in broader terms, a *construct* is *some postulated property or process characteristic of people, groups, organizations, situations, or environments that is assumed to be reflected in scores on measures of attributes of the property or elements of the process*. Constructs are important components of theories in social psychology (Shoemaker et al., 2004). Theories delineate constructs and posit the nature, direction, and limits of relations between them. Thus, because of their connection to constructs, measures can be profitably used to examine hypotheses and descriptive accounts put forth in social-psychological theories. These connections

between theory, delineation and measurement of constructs, and their use in research are depicted in Figure 2.

***Figure 2*: Theories Specify Constructs and Their Relations**



Note: Measurement is a primary means by which scores (x, y) are produced, enabling hypothesis testing (e.g., r) and description and informing theory.

Interplay Of Constructs And Measures

The properties or processes to which constructs refer differ in kind, and those differences have implications for how they are measured and modeled. A fundamental consideration is whether constructs are real. There are various interpretations of what the word "real" means. For some, the reality of constructs means that they serve relevant causal roles in relation to the measures (Borsboom, 2005); for others, whether their existence "is not mediated by human thought, language use, or empirical measurement" (Yarkoni, 2020, p. 328). Yet another way of understanding the reality of constructs is by requiring that they do not change as a function of how and when they are measured (Fried, 2017); in such a view, their properties are intrinsic and essential (i.e., they can be understood as natural kinds; Kripke, 1980). Examples may include constructs that refer to biological attributes or processes (e.g., amygdala activity and the social anxiety construct; Lieberz et al., 2022).

Measurement of constructs assumed to be real in this strong sense requires objective and accurate approaches. Constructs that are not natural kinds are, by their nature, subjective, and, as such, the accuracy of their measurement cannot be judged in the way that, for example, physical measures can, as one needs to consider the fact that they involve various social and cultural conventions and values. Such constructs are unequivocally the norm in social psychology. Sometimes, such constructs can be seen as useful metaphors (i.e., not to be taken literally) rather than natural kinds (Peters & Crutzen, 2017). Unlike natural kinds, which are discovered, these constructs are produced (Fried, 2017). Their meaning is a function of theoretical, social, and cultural interpretation; they are not merely a reflection or representation of the physical structure or the world (Messick, 1981). Examples include the "sociometer" of sociometer theory (Leary, 2012), the "resource" of ego-depletion (Baumeister et al., 2000), and the core social-psychological construct of "attitude" (Allport, 1935). Frequently, such constructs are socially constructed, existing as shared understanding or interpretation of regularities in people, groups, objects, and situations (e.g., Gergen, 2011). They may also reflect scientifically and practically useful entities that exist because they have predictive or interpretational value whether or not they are posited by theory. Other constructs are defined by

clusters of attributes that frequently co-occur due to causal connections that, together, reflect the construct (e.g., Schmittmann et al., 2013). Critically, and as detailed below, the nature of a construct has important implications for how it can or should be measured (Peters & Crutzen, 2017).

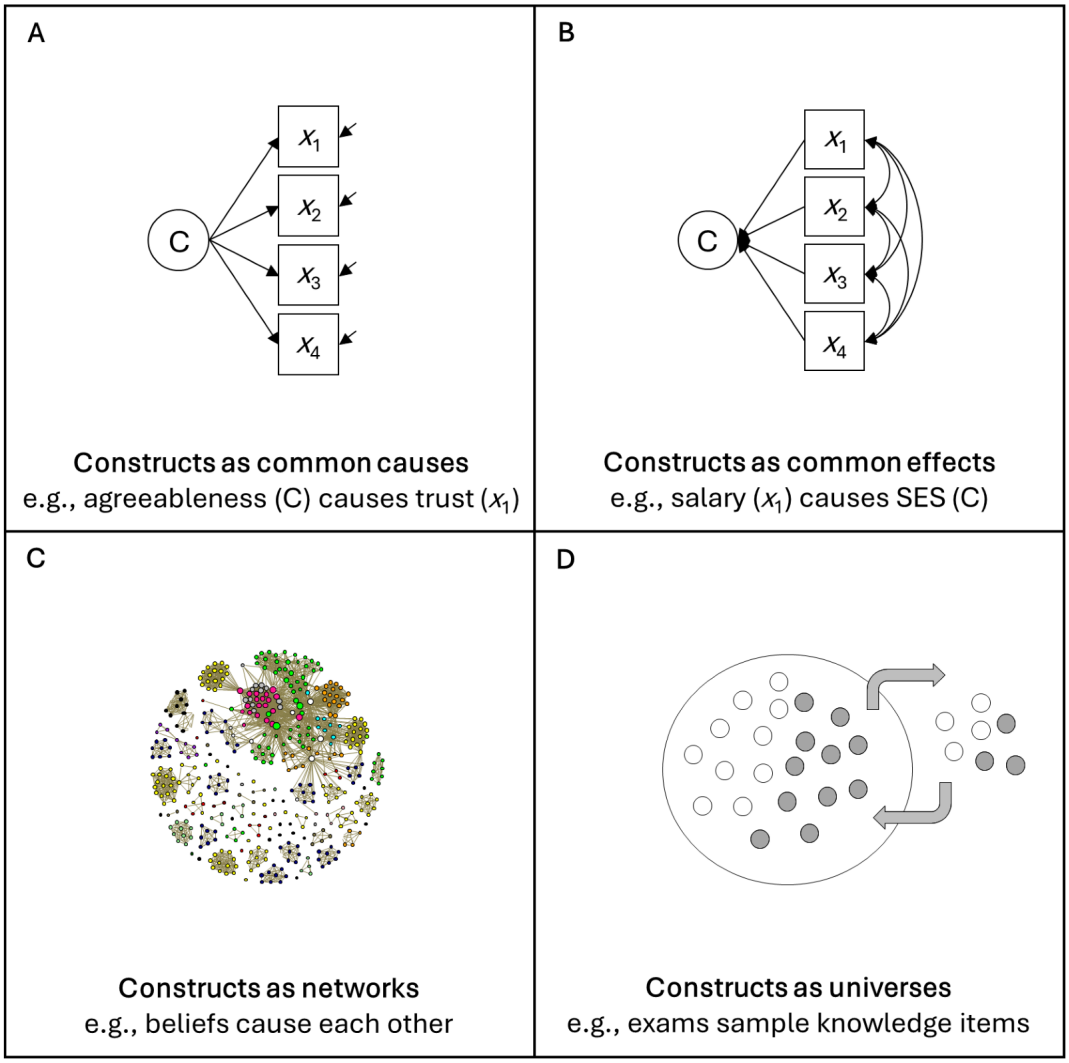
Several characteristics of social-psychological constructs and their interrelations are relevant for evaluating their quality and usefulness and for developing measures of them. Foremost among them is a lack of clarity and attention to differences between constructs in conceptualization and naming. This lack of clarity contributes to the well-documented jingle-jangle fallacy (Block, 2000; Weidman et al., 2017), which results from treating dissimilar constructs as if they are the same (*jingle*; e.g., referring to emotion regulation, empathic accuracy, and related constructs as socioemotional competence, Olderbak & Wilhelm, 2020) and similar constructs as if they are different (jangle; e.g., low agreeableness and "dark" personality, Vize et al., 2021). A related concern is clarity about the nature of the relations between constructs that are similar but distinct—*sibling constructs* (Lawson & Robins, 2021). Identifying sibling constructs—for example, self-esteem and narcissism (Hyatt et al., 2018)— and mapping out their conceptual similarities and differences aids in the development and evaluation of measures of them. A particular concern involving sibling constructs arises when scores from measures of each are contaminated by the other, resulting in ambiguity in interpretation and inference (Hoyle et al., 2023).

Achieving clarity about differences between constructs and their measures is an ongoing process, whereby conceptualizations and measures are updated in light of new evidence (e.g., results from using the construct or measure for research with a population or research question not previously considered) with the goal of increased clarity. In this process, a challenge is avoiding allegiance to an initial, perhaps influential, conceptualization when the goal is a maximally useful conceptualization that may require more than simply updating the initial conceptualization (Moreau & Wiebels, 2022). As conceptualizations of constructs are updated, their measures must also be updated, ensuring alignment between construct and measure (De Boeck et al., 2023). This process of improvement in construct specification and measurement is sometimes short-circuited when researchers in a topic area remain committed to an established measure despite refinements or updates to the specification of the target construct. The more scientifically sound alternative is a dynamic, cyclical process of improvement in construct specification and measurement with the ultimate goal of alignment between construct delineation, measurement approach, analysis model, and interpretation (Mari et al., 2021).

Statistical Representations Of Constructs

Two critical questions must be addressed in order to construct a psychometric theory for a given construct. First, how should the construct be characterized mathematically in a structure that reflects the associations between individual measurements and between those measurements and the construct? Second, how does that structure relate to available data? Answering these questions is tantamount to formulating a *construct theory*—a characterization of what a construct is and how it relates to empirical data collected to represent it. Four characteristic approaches to formulating a construct theory are represented in Figure 3. These do not exhaust the possibilities; however, they are approaches relevant for measuring many constructs in social psychology.

***Figure 3*: Four Ways in Which Constructs Can Relate to Data**



Note: In Panel A, the construct, C, is represented as the common cause of the indicators, x_1 to x_4; in Panel B, the construct is represented as the common effect of the indicators; in Panel C, the indicators are part of a network of interacting variables; in Panel D, the indicators are samples taken from a larger universe of behavioral elements.

First, a common way of thinking of psychological constructs is as a dimension. If the person's position on this dimension is considered a common determinant of multiple indicator variables, then it would be natural to represent it as a **common cause**, as in Panel A of Figure 3. A common cause screens off correlations between its effects, which in measurement models translates to the property of local independence. A wide variety of measurement models follow from this conceptualization. By varying the structure of the dimension (e.g., categorical or continuous, unidimensional or multidimensional) and the nature of the indicators (e.g., dichotomous, polytomous, continuous, etc.), one arrives at well-known psychometric models (Mellenbergh, 1994). For instance, in the case that the indicator variables are continuous, then the resulting model is the congeneric model (Jöreskog, 1971), better known as the single-factor model; if the indicators are dichotomous, the resulting model is an Item Response Theory (IRT) model such as that of Rasch (1960) or Birnbaum (1968). This conceptualization is commonly used in the assessment of personality. For example, extraversion is often thought of as the common cause of scores on measures of certain traits (e.g., gregariousness, assertiveness, excitement seeking), and can motivate the use of a factor model.

Second, there are cases in which the construct is conceptualized as a dimension, but, instead of a common cause, that dimension is a **common effect** of the indicator variables (Bollen & Lennox, 1991), as represented in Panel B of Figure 3. This representation may, for instance, be appropriate when a questionnaire asks for factors that are plausible effects of the construct in question. An example is the construct of life stress, which is typically assessed by asking whether a person has recently changed jobs, had a divorce, lost a loved one, and so on. Importantly, none of these questions directly query life stress; there is no reason to suspect responses to them would be correlated. Rather, they could be construed as causes of the construct (Bollen & Lennox,1991). In such cases, the relation between construct and observations can be specified as a formative model (Edwards & Bagozzi, 2000). In such a model, the correlation structure between the indicators is not directly modeled but is typically treated as a nuisance. Although the models in both Panels A and B can motivate the use of summed item scores, the interpretation of these scores is radically different. In Panel A, the total score is a measurement of a latent (i.e., inferred but not observed) psychological construct; in Panel B, the total score is an index, which tracks the movement of the indicator variables, much like a stock exchange index tracks the sales in underlying stocks.

Third, there are cases in which neither the common cause model nor the common effect model is appropriate. An interesting case arises if the indicator variables plausibly cause each other. For instance, in research on depression, symptoms that are typically queried include loss of sleep, fatigue, concentration problems, and worry. The correlations between these indicators may result from interactions between these symptoms, for example, insomnia → fatigue → concentration problems → worry → loss of sleep. In such a case, a plausible representation of the construct is a **network** (Borsboom & Cramer, 2013), which represents indicator variables as nodes and relations between them as connections between nodes. This representation is shown in Panel C of Figure 3. Typically, the relations among variables are estimated using partial correlations (Epskamp et al., 2018) or vector-autoregressive models in time series (Bringmann et al., 2013), but other methods are available. This representation has been argued to be appropriate for attitudes, because attitude elements (e.g., feelings, cognitions, and behaviors relating to an attitude object) plausibly interact (Dalege et al., 2016, 2017, 2018; Dalege & van der Maas, 2020).

Finally, there are cases in which the relation between indicators and construct is not plausibly thought of as being causal, but rather should be construed in terms of a **population-sampling relation**, as in Panel D of Figure 3. An example is a memory test that queries several unrelated memories, or a knowledge test that queries unrelated knowledge elements (e.g., a quiz). In these cases, one can think of the set of items as a sample from a larger population of knowledge elements (in other cases, these may also include behaviors or affective states). This population is often called a "universe" (Cronbach et al., 1972) or "domain" (McDonald, 2003). One can think of the construct score of any given person as the total score on the entire population of items, and of the test as a more or less random sample from that population. This means that the inference one makes from the scores on indicator variables to constructs is not a causal inference but a case of statistical generalization (Markus & Borsboom, 2013). Psychometric models that are associated with this way of thinking are those indicated as "strong true score theory" (Lord & Novick, 1968) and generalizability theory (Brennan, 2001).

Different construct theories may, but need not, motivate different psychometric models. As an example, consider the simple unidimensional IRT model. In this model, each of the dichotomous items in a set is related to a continuous latent variable through a monotonically increasing function (typically taken to be logistic). The model may be interpreted as a common cause model, wherein the latent dimension causes the item scores; a typical example is an educational test, in which

different items are included (e.g., addition items) requiring the same ability (e.g., arithmetic ability). However, the model can also be motivated by considering the construct to be a total score on an infinite number of items from which the items were sampled (Ellis & Junker, 1997). The model also follows from the hypothesis that the indicators form a fully connected network (Marsman et al., 2018). Thus, although construct theories and psychometric models are typically aligned, one has to be careful with overinterpreting the relations between the two.

It is important to note that even when different construct theories motivate the same model, it may nonetheless matter which one is chosen. Indeed, the interpretation of results is often very different under different construct theories, so it matters a great deal whether one, say, thinks of IQ scores as measures of a real latent ability (Borsboom et al., 2003) or as an index variable that summarizes the joint developmental trajectory of several elementary abilities (van der Maas et al., 2014). These interpretations suggest different kinds of interventions, different research lines, and different empirical tests. For example, how implicit and explicit features of the attitude construct are conceptualized and modeled (e.g., Perugini, 2005) affects whether and how each is accounted for in studies of the other (e.g., covariate, mediator, moderator) and the development of attitude change strategies (Rydell & McConnell, 2006).

Thinking more deeply about constructs is generally advisable because it results in a better appreciation of the degree to which specific psychometric concepts (e.g., internal consistency, unidimensionality) matter in a specific research situation. For example, in a common causal construct theory of attitudes, reliability reflects the degree to which indicator variables predict variance in the underlying dimension (the attitude). As a result, reliability is a normative concept, and it should be as high as possible. In contrast, however, in a network conception of attitudes, low reliability may be a feature of the attitude in question; for example, attitudes low in accessibility are characterized by a weakly connected network (Dalege & van der Maas (2020). This suggests that, in certain cases, low reliability as reflected in internal consistency values (e.g., alpha, omega) may not signal that there is something wrong with the set of items; rather, it may mean that one is assessing a weakly-connected network (additional issues in the interpretation of traditional reliability estimates are discussed in the Properties of Scores section later in the chapter.) Thus, the normative force of psychometric concepts can be significantly moderated by what construct theory is taken to be most likely for a particular measure and construct. The delineation and measurement connections shown in Figure 2 are the points in the research process at which these issues should be thoughtfully considered.

Latent Variable Models Of Constructs

In social psychology, the construct theory most often followed is reflected in *the latent variable hypothesis*. This hypothesis holds that a set of observed variables shares a common dependence on a latent structure. This common dependence is typically represented in a common cause model, which implies that the observed variables are associated because they all reflect the latent structure (Figure 3, Panel A).

Consider the following situation: three different thermometers are used to measure ambient temperature in the same room—a mercury thermometer, a digital thermometer, and people's subjective assessments. The readings on each instrument will rise and fall together as the temperature changes because they all depend on room temperature. However, because the readings do not explicitly affect one another, their correlations are spurious: controlling for the ambient

temperature will make the correlations disappear. This example illustrates the three cardinal assumptions of one of the most general psychometric models, the homogenous monotone IRT model, namely monotonicity, homogeneity, and local independence: (a) each of the thermometer's readings rises as a function of increasing temperature (monotonicity), (b) all of the thermometers only depend on the same latent variable namely temperature (homogeneity), and (c) given the position on that latent variable (i.e., if the temperature is held constant), the correlations between the thermometers vanish (local independence). The family of models covered by these assumptions is very large (Mellenbergh, 1994), and, for instance, includes the Rasch (1960) model, the two-parameter logistic model (Birnbaum, 1968), the congeneric one-factor model (Jöreskog, 1971), and a variety of nonparametric models (Sijtsma & Van der Ark, 2020).

Importantly, latent variables cannot be "constructed from the data" in the way that numerical representations of RMT can, because relations between people with respect to a latent variable are not directly accessible (Bollen & Hoyle, 2023; Borsboom, 2008). Instead, therefore, latent variables are typically elevated to the status of hypotheses. That is, they are hypothesized structures that determine patterns in the data (Bollen & Pearl, 2013; Edwards & Bagozzi, 2000). Therefore, measurement models that use latent variables should be considered "mini-theories" that incorporate these hypotheses. Because they are not in any way methodologically "given," they require substantive support. As a result, justification is required for representing the measurement process in the manner reflected in a *measurement model*, whether it is a factor model, IRT model, latent class model, or one of the many other models in use. Thus, one has to approach latent variable models as one approaches scientific theories—by testing the predictions and implications of the model. In this way, the measurement hypothesis is brought under scientific investigation in much the same way that other hypotheses are (Cronbach & Meehl, 1955): one is essentially investigating the veracity of the latent variable hypothesis by testing its empirical implications.

CONSEQUENCES OF INSUFFICIENT RIGOR IN MEASURING CONSTRUCTS

As illustrated in Figures 1 and 2, measurement is central and indispensable for research in social psychology. It is the bridge between constructs and empirical tests of hypotheses about them. For this reason, thoughtfulness and rigor in all other aspects of the research process (e.g., research design, sampling, statistical analysis) cannot overcome a lack of thoughtfulness and rigor in measuring constructs. Yet, perhaps more than any other aspect of the empirical research process, measurement is difficult. Rigorous measurement requires well-delineated constructs and measurement strategies that yield scores that can be attributed to those constructs. When it is achieved, the result is a relatively small gap in meaning between constructs and scores from measures of them (De Boeck et al., in press), which supports clear and meaningful inferences and interpretations.

Large or uncertain gaps in meaning between constructs and scores are an inevitable consequence of poorly delineated constructs and measurement strategies that do not reference and represent well-delineated constructs when they have been developed. In such cases, the informativeness of research that implicates the constructs is limited. The inferences, interpretation, and potential application of research—primary justifications for undertaking research—are uncertain and prone to error. In the column to the left in Table 1 are five specific ways in which the conceptualization and measurement of constructs could be insufficiently rigorous; though not exhaustive, the list is

illustrative. Associated with each instance of insufficient rigor is one or more specific consequences. Although the specific consequences vary across instances of insufficient rigor in measuring constructs, they have in common the effect of reducing the overall rigor and value of research studies. The implication is that each instance of insufficient rigor described in the table would need to be addressed to ensure rigorous research that fully serves its purpose.

***Table 1*: Examples of Insufficient Rigor in Measuring Constructs and Their Consequences**

Insufficiency of Rigor	Consequences
The construct of interest is not sufficiently well-delineated to inform the measurement of it.	The relevance of results from research using measures assumed to reflect the construct is uncertain. Hypotheses about the construct are imprecise and potentially unfalsifiable.
The chosen measurement strategy is not intentional in identifying and choosing the target (e.g., self, group, situation), manifestation, and source of information about the construct.	The correspondence between scores and the construct, as delineated, is unclear. Inferences based on scores may not account for potential sources of bias or missing attributes.
Measures of a construct of interest do not capture its primary attributes as specified in a well-delineated conceptualization.	Research results based on scores from a measure may underestimate or fail to detect associations involving the construct because scores do not fully reflect it.
How scores are produced by the measurement strategy is not sufficiently clear or justified.	How scores and differences between them are interpreted is unclear. How different attributes of the construct are reflected in scores might be unclear. These issues are pronounced for scores from machine learning algorithms used to generate scores from organic data.
A body of research on a particular topic relies on a single measure of the focal construct.	No measure or measurement approach perfectly captures the targeted construct. A body of work based on a single measure is limited by the flaws of that measure. Shortcomings of the measure limit the quality and impact of research on the topic.

Rigorous measurement of constructs in social-psychological research can take many forms. These vary along a set of dimensions that define a framework for discussing the breadth and diversity of conceptualizations and measurement strategies in social psychology. A consideration of these dimensions highlights key theoretical and practical considerations in delineating and measuring constructs with rigor and creativity.

A FRAMEWORK FOR BREADTH AND DIVERSITY IN MEASUREMENT

The diverse array of constructs of interest to social psychologists necessitates diversity and creativity in their measurement. This section describes an organizing framework as a context for discussing the options and opportunities for conceptualizing and measuring constructs in social psychology. The framework, summarized in Table 2, organizes the broad array of constructs to be measured according to who or what they concern (i.e., targets), how variability in them may be evident (i.e., attributes), and the ways they may be expressed (i.e., forms). A variety of sources of information from which measures could be derived are identified. These elements—targets, attributes, forms, and sources—and their many combinations identify a broad and diverse array of constructs and measurement strategies.

Table 2: *Elements of a Broad Perspective on Conceptualizing and Measuring Constructs in Social Psychology*

Targets	Attributes	Forms	Sources

• people
  ○ individuals
  ○ dyads
  ○ groups
  ○ institutions
  ○ cultures
• environments
  ○ objects
  ○ situations
  ○ virtual experience
	• behavior
	  ○ observable
	  ○ unobservable
	• cognition
	• emotion
	• biology
		• consistency
		  ○ trait
		  ○ state
		• breadth
		  ○ general
		  ○ specific
		• awareness
		  ○ explicit
		  ○ implicit
			• self
			• informants
			• specialized equipment
			  ○ equipment not developed for research
			  ○ ambulatory assessment
			  ○ passive sensing
			• organic data
			  ○ social media
			  ○ Internet of Behavior

Targets Of Conceptualization And Their Measurement

The constructs specified by theories and manipulated or measured by social psychologists can be differentiated according to the aspect of social experience, or target, they concern. Principal among those targets is the person. The primacy of this target in measurement by social psychologists is evident in the many measures and measurement strategies designed to detect variability in people's perceptions, experiences, preferences, intentions, and so on. Though predominantly focused on the individual, measurement may target collections of people as in dyads, groups, families, and organizations. The thoughts, feelings, and actions that constitute social experience occur in places and spaces that constrain and shape the experience. These potential targets of measurement include both objective environments (e.g., outdoors, in the classroom) and subjective situations, which may vary within an environment (e.g., evaluation, normative pressure). Although these aspects of social experience are frequently manipulated, they can be measured as well. Certain measurement challenges and opportunities are common to all constructs associated with a target type and warrant consideration when selecting a measurement strategy or approach.

People

Most measurement undertaken by social psychologists characterizes individuals, about whom inferences will be made in research using scores. A significant benefit of individuals as targets is

their ability to report on themselves, which is reflected in social psychology's preference for self-report measurement. However, as detailed below, well-delineated constructs that refer to individuals may be measured using other information sources, some better suited for certain attributes specified by constructs.

For constructs that characterize the shared experience of two or more people, the potential sources of information are less clear. For example, measurement of the construct of interdependence in dyads might involve combining individual self-reports of interdependence in the relationship, or it could involve third-party observations by acquaintances, trained judges, or experimenters. The use of self-reports from individuals to characterize the standing of a collection of people on a construct may not be feasible as the number of people and the social or physical distance between them increases. Examples range from small groups of a fixed size composed of a constant set of people who interact frequently, to larger groups or groups for which size and membership are not fixed. The choice of measurement strategy in these instances is dictated by two related concerns—the construct of interest and the *unit of analysis*. Constructs that refer to the characteristics of the group assume that the unit of analysis is the group. Alternatively, constructs may refer to individuals' experience of belonging to the group and assume that the unit of analysis is the individual (e.g., perceived cohesion; Bollen & Hoyle, 1990). In rare instances when the group is the unit of analysis, it might be feasible, as with dyads, to combine group members' ratings of the group on attributes specified by the focal construct to produce a single score (Chan, 1998). In most instances, however, it will be necessary to use alternative approaches that take advantage of existing information about the group (e.g., racial composition, performance history) or acquire third-party ratings from observers.

In terms of number and diffuseness, the most extreme versions of measurement focused on collections of people are those that target large organizations and nations or cultures. As with smaller collections of people, the construct and unit of analysis may be focused at the individual level; however, because it is not feasible to collect data from all members of large organizations or cultures, representative samples are required (e.g., Rentfrow, 2010). If the construct and unit of analysis are not the individual, as in comparisons of organizations or countries, then metrics that characterize the whole (e.g., average salary, Gini Index) are potential measures of attributes specified in the construct. As is evident across instances of people as the target of measurement, how measurement is done is intimately tied to the unit of analysis, which follows from the nature of the construct about which inferences will be made.

Environments

Although most measurement in social psychology focuses on constructs that characterize people, it may concern other targets. Specifically, the focus of measurement may be constructs that characterize situations, environments, and objects other than people in them (Block & Block, 1981). From a measurement perspective, the simplest case is constructs that specify attributes of people's perceptions of or preferences for situations, environments, and objects. In such cases, self-reports and other expressions of perception or preference are suitable (e.g., Funder et al., 2012). If, however, the construct of interest concerns objective features of these targets that differentiate between them (e.g., Kelly et al., 2003; Reis, 2008), then alternative approaches are required. These might include objective ratings of specific environments such as those used to characterize neighborhoods using images captured by Google Street View (e.g., Odgers et al., 2012; see Andresen et al., 2013, for an alternative approach), specific objective features of situations such as their duration and

concreteness (Rauthmann & Sherman, 2021), or the number of people present, their proximity to each other, and their social status (Latané, 1981; Nowak et al., 1990). Objects in environments are potentially relevant for social experience in their own right (e.g., the presence of a gun or mobile phone). Measuring their physical properties (e.g., make, model) may be quite straightforward; however, they may also be viewed as part of certain environments that, together with other features of the environment, constitute *social space*, which can be conceptualized and measured at different levels of analysis (e.g., Balsa-Barreiro et al., 2022; Schafer & Schiller, 2018; for a review, see Martin et al., 2017).

A relatively new and increasingly important manifestation of spaces and places is the **virtual environment** (Boczkowski & Mitchelstein, 2021). Of potential interest to social psychologists is the degree to which constructs relevant to social experience outside virtual environments apply to social experience within them. In terms of measurement, an intriguing feature of the virtual environment is the ability to characterize it using information from the same digital devices that enable it (Bouwman et al., 2013; Harari et al., 2016). Moreover, those devices may be used for active (and intrusive) measurement typical of social psychology (e.g., daily self-reports of digital experiences; Dwyer et al., 2018) or passive measurement using sensors (e.g., location, accelerometer) and apps that unobtrusively capture information about features of social experience (e.g., Stachl et al., 2020). The opportunities afforded by digital devices may enable the measurement of attributes of the spaces and places in which social life is enacted that have received little empirical attention by social psychologists outside of laboratory experiments.

Attributes Of Constructs And Their Measurement

Constructs that apply to people may refer to different types of psychological attributes, such as biology, emotion, cognition, or behavior. Although some constructs, by their nature, refer to a single type of attribute (e.g., physical activity, person perception), most encompass multiple types of attributes (e.g., attitude, self-regulation, well-being). The attributes specified by construct delineations determine the appropriateness of different approaches to measuring them. Well-delineated constructs also specify the theoretical connections between attributes.

Behavior

Some constructs concern, entirely or in part, what people do. Although behavior is, in theory, always observable, some behaviors are, practically and ethically, not subject to observation for research purposes. Thus, when measurement concerns constructs that include behavioral attributes, the choice of measurement approach must first consider whether the behavior can be observed. If observation is practically achievable, then the concern is how to do so without altering the behavior (i.e., the Hawthorne Effect; McCambridge et al., 2014). If, for practical or ethical reasons, the behavior cannot be observed, then the concern is how to secure accurate reports of it.

When observing behavior to measure its characteristics, the primary concern is how observation will be accomplished. The most straightforward option is to train observers or enlist people who routinely observe the behavior of members of the population of interest to detect evidence of the characteristics of interest and assign scores using a formal, preferably validated, rating system (e.g., Stevens et al., 2010; Vernham et al., 2016). Their observation may be the basis for ratings close in time to enactment of the behaviors or ratings that summarize observation of the behaviors over

time. It may be possible for observers to rate behavior covertly so that the research participants are not aware they are being observed (e.g., two-way mirror, hidden camera streaming video). Behavior may also be captured on video for later coding (as opposed to streaming and rating in real-time). Using data from covert observation and recording of which participants are unaware poses significant ethical challenges that must be addressed before their use (Podschuweit, 2024).

For practical or ethical reasons, the measurement of some behaviors by observing targets in real-time or through the use of video recording is substantially limited. Examples include the use of illicit substances, engagement in violent or destructive behavior, and sexual activity. The typical approach to measuring constructs that refer to behaviors of this type is self-report, by which research participants are both the behaver and the reporter of it. Self-reports of behavior, in general, are subject to several well-documented biases (for a review, see Haeffel & Howard, 2010). Accuracy is an additional concern for private or illegal behaviors not subject to observation (e.g., Harrison, 1995; Turner et al., 1997). Additional behaviors for which observation is not feasible or appropriate are behaviors that are not sensitive but cannot practically be observed or accurately self-reported. One example is digital device and social media use, for which—owing to their frequency and ambiguity in the definition of a use episode—self-reports are so inaccurate as to be useless (Parry et al., 2021). Although not perfect, **passive approaches** that use devices to capture and transmit digital behavior via sensors and apps are an option. Combining these passive approaches with active, device-enabled data collection can both reduce participant burden and increase accuracy in the measurement of certain behaviors (Davidson et al., 2022; Keusch et al., 2021).

Cognition

Some constructs refer exclusively to cognitive processes or include one or more cognitive processes among their attributes. Observation is not a viable option for direct measurement of cognitive processes apart from imprecise detection and labeling of activity in the brain (cf. Liu et al., 2022; Wallis, 2018). Moreover, self-reports of cognitive processes that contribute to behavior and decision-making are often invalid, especially for higher-order cognitive processes (Nisbett & Wilson, 1977; cf. Petitmengin et al., 2013). Nonetheless, self-reports of cognitive processing are frequently used as a measure of constructs that implicate those processes (e.g., Baksh et al., 2018; Farias et al., 2008). Although self-reports of cognitive processing are unlikely to reliably reflect the construct they were designed to measure (e.g., Craig et al., 2020; Herreen & Zajac, 2017; Klein et al., 2022), they may nonetheless serve as suitable measures of the presence or absence of skills that imply cognitive processing (e.g., Monahan et al., 2014; Rast et al., 2009).

Alternatively, cognitive processing may be inferred from performance on tasks or neurobiological activity while performing them. These indicators may be used to measure individual differences in cognitive abilities or deficits (e.g., Basile & Toplak, 2015) or to measure online cognitive processing in response to specific features of tasks (e.g., Sharifian et al., 2021). In terms of performance, **task-based indicators** of processing include metrics such as accuracy or correctness and speed (e.g., Hirsch et al., 2018). Indicators based on neurobiological activity during task performance include metrics such as pupil size and frequency of rapid pupil dilation (e.g., Vogels et al., 2018), and different signals of brain activity such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI; for a review, see Nichols & Newsome, 1999). A challenge with task-based measures of cognitive activity and skill is the general lack of association between these measures and self-report measures (e.g., Snyder et al., 2021) and ambiguity regarding the specific construct that is on display during the performance of a particular task or activity. For example, neural activity associated with

social cognition is frequently examined as research participants observe social interaction, which may differ from the neural activity and the construct it represents when they engage in social interaction (Schilbach, 2014).

Emotion

A third type of attribute relevant to conceptualization and measurement is emotion. Emotion constructs may focus solely on experienced emotion; however, emotion is implicated in many constructs primarily defined by cognitive or behavioral attributes. In terms of experienced emotion, constructs may refer to specific emotions, such as anxiety or fear, or features of emotional experience, such as valence or arousal. A particular challenge at this level of conceptualization and measurement is labeling, both by investigators (Weidman et al., 2017) and, in the case of self-reports, research participants. In terms of investigator labeling, the frequent disconnect is between measurement and a well-delineated emotion construct specified by a taxonomy or theory; for example, measures imply more emotions than taxonomies specify (Weidman et al., 2017). In terms of labeling by research participants, common language use may refer to emotional experience in ways that do not map well into emotion space as specified by taxonomies or theories (Cowen & Keltner, 2017). For these reasons, the challenges in measuring emotion may result more from issues of construct specification than the specifics of measurement strategy or approach.

A critical distinction in emotion measurement (though not necessarily conceptualization) is between trait and state forms of constructs. Concerning measurement, the earliest instance of this distinction is the widely used State-Trait Anxiety Inventory (Spielberger et al., 1983). With the increase in experience sampling methods, the distinction is now more likely to appear as a baseline measurement of typical emotional experience (i.e., **trait**) and in situ measurements of momentary emotional experience (i.e., **states**). The former is assumed to be cross-situationally consistent. In contrast, the latter is expected to vary due to variability in situations and personal experience (Fleeson & Law, 2015). Using these methods, the extent of variability in the experience of specific emotions may be measured (e.g., Liu et al., 2019) and, in the interplay between construct and measurement, may lead to refinement in or expansion of the conceptualization of certain emotion constructs.

The approaches to measuring emotion constructs are perhaps more extensive and varied than for any of the attributes relevant for conceptualizing and measuring constructs. Because emotional experience is inherently subjective, self-reports are, for many emotion constructs, the preferred measurement strategy, whether the focus is a trait (e.g., Watson & Walker, 1996) or state (e.g., Harmon-Jones et al., 2016). Because reporting on emotional experience may affect it, frequent or continuous measurement requires alternative approaches, such as ecological momentary assessment (EMA), or an instrument that does not require effort and attention such as the affect rating dial (Ruef & Levenson, 2007). However, those types of measures are limited in the amount of detail they can provide on the emotional experience. Additional approaches include implicit measurement of both trait and state affect (Quirin & Bode, 2014), peer ratings (e.g., Watson & Clark, 1991), and an array of measures based on physiological activity (Mauss & Robinson, 2009). As with measures of cognitive constructs, scores generated by these different approaches do not always converge (Mauss et al., 2005), suggesting that either they tap different forms of the targeted emotion construct or differ in their effectiveness at capturing it. Beyond measures and markers of discrete emotions and their properties are numerous measures focused on emotion-related constructs, such

as emotion regulation, emotional intelligence, and broader constructs to which emotion is central (e.g., attitudes, prejudice, aggression, well-being).

Biology

Although biological attributes and processes are not inherently social or psychological, they are implicated in many social-psychological constructs (e.g., Bartels et al., 2008; Kitayama & Uskal, 2011). For example, the profound biological shifts during puberty result in hormonal and bodily changes with significant implications for social processes, such as interpersonal attraction, gender-related processes, and sexual behavior (Downs, 1990). The short-term and persistent neurobiological effects of substances, such as alcohol, amphetamines, and hallucinogens influence basic contributors to social experience, such as perception, inhibition, and decision-making (e.g., Steele et al., 1985). Genetic and emergent epigenetic influences account for a portion of the variance in a wide variety of social behaviors through their effects on physiology and the brain (Robinson et al., 2008; Seebacher & Krause, 2019). For these reasons, biological attributes and processes may be implicated in the conceptualization of a construct and, therefore, its measurement.

The measurement of biological attributes and processes differs from measurement of other attribute types because it may allow a level of precision and reference to objective units that could never be achieved for cognition or emotion. Yet, because the biological factors that contribute to social experience are the same factors that control routine bodily functions, they are, at best, imperfect indicators of the social and psychological factors for which they are relevant (in this respect it is also useful to note that, despite their objective character, such measures are often contaminated by significant amounts of noise). Biomarker assays from saliva, hair, blood, and other biological samples can serve as measures of behavior, social experience, and well-being (for a review, see Ewbank, 2008). Genetic profiles and polygenic scores may characterize risk or potential related to social-psychological constructs, though they do not always map onto constructs specified in theories or biological processes that confer risk or potential (Burt, 2022). Some aspects of biology are of interest to social psychologists due to the effects of social experience on biological structures and processes, such as, for example, in epigenetics (Champagne, 2010) and immunology (Shattuck, 2021). Collectively, these neurobiological attributes and processes are the focus of constructs in social neuroscience (Cacioppo et al., 2010) and social psychoneuroimmunology (Muscatell, 2021).

Forms Of Constructs And Their Measurement

In addition to target and attributes, constructs can be characterized by the forms of their expression and how, if at all, their expression differs across time and place. The most fully delineated constructs specify the type and degree of variability in the forms of their expression. Construct form varies along three primary dimensions. *Consistency* concerns the degree to which a construct and scores from measures of it are the same or vary at different points in time. *Breadth* concerns the degree to which a construct and scores are expected to be the same or vary across different tasks, activities, or life domains. *Awareness* concerns the degree to which people are conscious of and able to report on the presence and influence of the construct. The primary attributes of some constructs do not vary for different combinations of consistency and breadth, only their expression. For example, trait, state, general, and domain-specific forms of self-esteem measurement typically assume the same conceptualization and use straightforwardly adapted measures of the construct (e.g., Kernis et al., 1993; Marsh, 1986). The primary attributes of other constructs vary across forms of

expression (e.g., self-efficacy vs. generalized self-efficacy beliefs, Bandura, 2006; Schwarzer & Jerusalem, 1995). Nearly all constructs that may be expressed with and without awareness (e.g., implicit vs. explicit self-esteem, Pelham et al., 2005) differ in conceptualization as reflected in different measurement strategies at different levels of awareness. These similarities and differences in how constructs may be expressed have significant implications for how they are measured.

Consistency

Consistency refers to the degree to which evidence of the construct, as reflected in scores on measures of it, is constant or variable across time and place. Constructs that are relatively constant are commonly referred to as traits, and those that are variable are referred to as states (e.g., the experience of emotions described in the attributes section). In reality, whether delineated as such or not, many constructs are somewhere between the two (Geiser et al., 2017). That is to say, the attributes and processes specified by the construct have a stable form that is cross-temporally and cross-situationally consistent and a variable form that is sensitive to differences in internal and situational states. When the delineation of the construct is silent concerning stability and variability, statistical models that allow for both components of variance in scores on a measure across time or situations can provide insight into consistency of expression. For example, an application of the integrated trait-state model to personality ratings on 90 consecutive days found considerable intraindividual variability (i.e., "state" variance) in extraversion and neuroticism ratings (Hamaker et al., 2007). The model also revealed that the degree of variability across time varied between persons, suggesting that delineations of these constructs, typically considered traits, should account for day-to-day variability.

Evidence of constant and variable forms of constructs is influenced by how they are measured. For example, self-report items that refer to "in general" or "typically" invite ratings that are trait-like. Alternatively, items that refer to "right now" or "currently" invite ratings that are state-like. Importantly, however, statistical approaches such as the integrated trait-state model can reveal variability in trait ratings and constancy in state ratings. When accounts of constructs do not address consistency, the results of these models can lead to fuller and more informative delineation of constructs. For construct delineations that address consistency of expression, such as anxiety (Spielberger et al., 1983) and self-esteem (Kernis et al., 1993), through different versions of measures, these models offer rigorous tests of theoretical assumptions about the consistency with which the construct is expressed.

Breadth

Breadth concerns the degree to which constructs express similarly across tasks, activities, and spheres of life or vary in their expression across those domains. *Domain-general constructs* express similarly across domains, whereas *domain-specific constructs* may express differently in different domains. As with consistency, delineations of constructs might not address the breadth of expression, thereby leaving the determination of the degree of breadth of expression of constructs to empirical strategies. Those strategies typically entail the tailoring of a domain-general measure to focus on one or more specific domains. Administering the domain-general and one or more domain-specific versions of the measure allows for the estimation of similarity in scores across domains and the relative performance of domain-general and domain-specific scores in predicting outcomes of interest. For example, in examining the potential domain-specificity of self-control, Haws et al. (2016) tailored a standard measure of domain-general self-control to measure self-control in the

spending and eating domains. They found moderate correlations between scores on the domain-general and domain-specific measures and that correlations between domain-specific self-control and domain-relevant outcomes were stronger than correlations between domain-general self-control and the same outcomes (see Cronbach & Gleser, 1957, on bandwidth versus fidelity in measuring constructs for purposes of prediction). A similar strategy has shown a comparable pattern for measures of other constructs, including grit (Rumbold et al., 2022), risk attitudes (Weber et al., 2002), and quality of life (Wu & Yao, 2007). An alternative approach examines statistical models of the relations between domain-general and domain-specific expressions, often finding a hierarchical factor structure in which a correlated set of domain-specific factors are influenced by a second-order factor assumed to be the domain-general expression of the construct (e.g., Marsh, 1990). Results of these empirical considerations of breadth suggest that descriptions of many constructs in social psychology should account for potential variability in expression across tasks, activities, and life-domains.

Awareness

Awareness concerns the degree to which the target, nearly always a person, is conscious of the attributes and processes specified by a construct as it influences their emotions, cognition, and behavior. Explicit forms of constructs specify attributes and processes of which the person is aware. Conversely, implicit forms of constructs specify attributes and processes that exist and exert influence without awareness (for a review, see Nosek et al., 2011), though the person may become aware of their existence and influence on reflection (Hahn et al., 2014). As with differences in consistency and breadth, awareness of the expression of a construct is not fully present or absent. Rather, many constructs of interest to social psychologists may be expressed with or without awareness. Examples include prejudice (Brauer et al., 2000), self-esteem (Spalding & Hardin, 1999), and aggressive tendencies (Frost et al., 2007). Unlike the strategies that allow for comparable measurement of constructs across levels of consistency and breadth of expression, approaches to measuring explicit and implicit expressions of constructs are, by necessity, quite different. Whereas the measurement of explicit expressions of constructs is relatively straightforward, typically as self-reports, the measurement of implicit expressions is challenging. The dominant approach in the field is the Implicit Association Test (Greenwald et al., 1998; cf, Blanton et al., 2006), though other strategies have been developed (Jones et al., 2002; Payne et al., 2005). Not all constructs of interest to social psychologists express beneath awareness; however, for the many that do, fully-specified construct delineations require an accounting for explicit and implicit forms, the association between them, and when each is expected to influence emotions, cognition, and behavior (e.g., Payne et al., 2017). Measures of these two forms of constructs that are at least somewhat parallel would allow for statistical modeling that accounts for both forms as with models of differences in expression in terms of consistency and breadth.

Sources Of Information About Constructs

Different sources of information about a construct may be available for different combinations of target, attributes, and forms of measurement. Key considerations are minimizing bias and maximizing correspondence between the information available from the source and the construct. In terms of bias, the concern is the accuracy of information about specific behaviors (e.g., frequency or amount of alcohol use, time spent using social media apps) and validity of information about cognitive activity and emotional experience, especially when the source is subject to motives that

influence reporting (e.g., underreporting undesirable emotional experiences) or unable to access relevant information (e.g., self-reports of higher-order cognitive processes, observer reports of emotional expressiveness). For these reasons, any particular source of information may be inadequate for measuring a construct, pointing to the value of triangulating constructs in measurement by obtaining construct-related information from multiple, diverse sources (Eid & Diener, 2006; Smith & Harris, 2006; for an example, see Connors et al., 2016). So doing permits the separation of method and construct variance in scores produced by different measurement strategies. For example, four sources of information from which measures of social-psychological constructs could be drawn are the self, informants, specialized equipment, and organic data. Scores from each source, assumed to share only construct variance, should relate similarly to other constructs (Campbell & Fiske, 1959).

Self

Social psychologists are more reliant than ever on research participants as the primary, even sole, source of information about their own thoughts, feelings, motives, and behavior (Sassenberg & Ditrich, 2019). The foremost consideration when considering the self as the source of information about a construct is whether, for that construct, the participant is able and willing to report on the attributes it specifies. Ability concerns include whether participants have access to the attributes or processes of interest (Nisbett & Wilson, 1977) and whether they have reliable recall of prior experiences, states, and behaviors (e.g., Larsen & Fredrickson, 1999; Verbiej et al., 2021). Assuming ability, willingness is evident in the influence of personal motives and concerns (e.g., social desirability) that result in reports that do not reflect the participants' true standing on a construct (e.g., Brenner & DeLamater, 2016). Other considerations include context (e.g., Hadwin et al., 2001), modality (e.g., Gnambs & Kaspar, 2015), and reference standard (e.g., Lira et al., 2022), which is addressed in the section on self-report measurement.

Informants

For certain constructs, particularly those for which the self is unsuitable as the sole source of information, informants may be a viable source. The term informant broadly refers to any person who provides information about a target's standing on a construct. Although the targets of informant measurement typically are other people, they could be objects, situations, or environments. A key concern is the degree to which the attributes of interest are observable or can be inferred with validity via observable actions or characteristics (Carpenter et al., 2017; John & Robins, 1993). A related concern is whether the informant is sufficiently informed and attentive to provide valid ratings of the target on the construct of interest. The choice of the informant must consider the characteristics of the target to be rated and potential informants' exposure to the target, accepting that a single informant perspective may be insufficient to fully capture the construct (e.g., Coie & Dodge, 1988). The relationship between the informant and the target is relevant for many constructs and can range from people unfamiliar with the target (naïve or expert; e.g., Gosling et al., 1998) to people familiar with it. When the target is a person, familiar informants include peers (e.g., Larson et al., 2007), coworkers (e.g., Heainisch & Jex, 1998), roommates (Paunonen & Kam, 2014), romantic partners (e.g., Foltz et al., 1997), and, for younger targets, parents and teachers (e.g., Warnes et al., 2005). When the target is an object, situation, or environment, knowledge about the target through experience or training is desirable if not essential (e.g., Andresen et al., 2013; Linney, 2000).

A special case of informant-based measurement involves each member of a collection of people rating themselves and each additional person in the set (Kenny & La Voie, 1984). The set may include members connected in some meaningful ways, such as families (Snijders & Kenny, 1999), work teams (Kluger et al., 2021), and other intact groups (e.g., Øverup et al., 2021), or members who are previously unacquainted (e.g., Dufner & Krause, 2023). The full set of self- and other ratings by all people in a set allows for the decomposition of variance in ratings on a construct into components corresponding to how people, on average, rate others on the construct; how people, on average, are rated on the construct by others; and how people are rated differently on the construct by specific others. The round-robin designs typical of applications of the Social Relations Model demonstrate the value of informant ratings for understanding the sources of variance in a construct. That value is also evident in the use of collateral reports to triangulate targets' true standing on a construct (e.g., Connors & Maisto, 2003).

Specialized Equipment

Information about some constructs may be obtained from specialized equipment either designed specifically for psychological measurement or for general use but nonetheless a source of construct-relevant information. When the conceptualization includes biological attributes, specialized equipment may be the only viable source of information. For other attribute types, it may be used to complement information from other sources, such as self-reports or observer ratings. And, for some constructs, specialized equipment may be the only option for measuring some attributes (e.g., response latency).

Focusing first on constructs conceptualized primarily in biological terms, when attributes of the construct involve brain activity, a widely used approach to measurement is fMRI using a scanner (e.g., Zarate, 2023). Scanners are examples of general-purpose equipment that produce output useful for measuring constructs of interest to social psychologists (though how those measures are extracted is a concern; e.g., Eklund et al., 2016). Although resting-state fMRI scans are useful for characterizing some attributes or as an informative comparison (e.g., Pang et al., 2022), fMRI scans typically measure activity in selected brain regions while the participant in engaged in a task assumed to require the cognitive or affective process of interest (for a review, see Phua & Christopoulos, 2014). Other uses of specialized equipment for measuring brain activity include electroencephalography (e.g., Harrewijn et al., 2018) and magnetoencephalography (e.g., Levy et al., 2016) and the associated event-related potential and event-related field patterns of activity, respectively (e.g., Schmid & Amodio, 2021).

Moving beyond brain activity, several biological signals provide social psychologists with information about constructs of interest. Examples include skin conductance (Boucsein et al., 2012), heart rate and heart rate variability (Berntson et al., 1997), blood pressure (Shapiro et al., 1996), and pupillometry (Mathôt, 2018; for a review, see Cacioppo et al., 1989). Particularly promising for situated and continuous measurement is equipment that enables ambulatory assessment, which, when paired with EMA and passive sensing by mobile devices, allows for highly informative capture of information relevant to multiple attribute types (Conner & Mehl, 2015; Trull & Ebner-Priemer, 2014). Although many fixed and ambulatory systems for capturing physiological information are designed specifically for research, the information they provide, often in a continuous stream, requires substantial processing. Strategies for evaluating measures derived from these sources are more challenging and less frequently applied than strategies for evaluating measures derived from other sources, such as self-report or informant ratings (Strube & Newman,

2007; Tomarken, 1995). Yet, they have the potential to provide information about constructs that either cannot be accessed by self or observers or are subject to biases that limit the usefulness of information from those sources.

Perhaps the most widely used specialized equipment in social psychology is the computer. Moreover, a significant proportion of social interaction is now accomplished using computers and other Internet-connected devices. A feature of the computer that has proven particularly useful in measuring constructs is the ability to record with precision the amount of time between keystrokes. Keystrokes paired with stimuli presented on the computer screen (super- or subliminally) can be used to measure response latency, a useful source of information about attention, deliberation, processing speed, and other attributes specified by constructs (Fazio, 1990). Examples include attitude accessibility (e.g., Mulligan et al., 2003), implicit cognition (e.g., Greenwald et al., 1998), emotional arousal (e.g., Temple, & Geisinger, 1990), and implicit emotional states (Bartoszek, & Cervone, 2022). As with scores from physiological measurements, the interpretation of scores from computer-based measures such as response latency is not straightforward (for an example, see Blanton et al., 2009).

Most uses of specialized equipment as a source of information about a construct are done in such a way that research participants are aware that something about them is being measured. An alternative approach makes use of **passive sensing**, often via electronic devices already in use by participants or wearable devices that are soon forgotten by the wearer (for a review, see Schödel & Mehl, 2025). The approach is passive in that, once measurement has begun, the research participant is not required to engage with the device for the purpose of measurement. It detects different forms of activity through sensors native to the device, typically a smartphone (Cornet & Holden, 2018; Stachl et al., 2020). Within psychology, popular forms of passive sensing capture individual behaviors, such as location (e.g., GPS), movement (e.g., accelerometer), health status (e.g., physiological sensors); or relational behaviors, such as interactions with others (e.g., sociometric badges) and online communication (e.g., social media; see Saha et al., 2019; Thapa et al., 2021; Wu et al., 2008). In addition to smartphones, devices useful for passive sensing include smart watches (Thun et al., 2012), wrist and ankle bracelets (e.g., Greenfield et al., 2014; Sultana et al., 2018), smart rings (e.g., Moshe et al., 2021), wearable cameras (Brown et al., 2017), and smart-home technology (Nelson & Allen, 2018). Although there are many technical aspects of passive sensing, such as signal processing and backend data management (for reviews, see Bayoumy et al., 2021; Harari et al., 2016), the approach has considerable promise as a means of unobtrusive measurement in situ.

The promise of rich and continuous data passively collected in natural settings is offset somewhat by strategic and technical concerns. Using novel devices may require acclimation and may create irregular behavioral repertoires. For example, wearable cameras that take snapshots throughout the day need participants to temporarily remove them when bystanders are uncomfortable or in non-public spaces (Brown et al., 2017). Wearing an ankle alcohol monitor may result in less naturally occurring alcohol consumption behavior (Greenfield et al., 2014). Furthermore, because devices may not be reliable, careful validation and pilot testing may be required to determine their viability for a specific context (e.g., Chaffin et al., 2017; Greenfield et al., 2014). More broadly, in research that extracts information from devices purchased and worn outside the research context (e.g., Fitbit, Apple Watch, Oura Ring), the non-representativeness of people using such devices may increase inequities in research (Bayoumy et al., 2021) and, therefore, the generalizability of findings. Finally, a number of ethical considerations related to privacy and data security beyond those typical of research in social psychology require careful consideration (Kreuter et al., 2020).

Organic Data

Organic data are a wide-ranging information source increasingly available for behavioral research and of potential value for measuring constructs of interest to social psychologists (see Kosinski, 2024). Unlike designed data generated by researcher-managed sources, organic data are generated as digital byproducts of people's interactions, transactions, and movements (Groves, 2011). The volume of organic data is astounding and growing. As an example, in one minute across the globe in 2019, more than 18 million text messages were sent, 188 million email messages were sent, nearly 4 million Google search queries were entered, and 4.5 million YouTube videos were viewed (Desjardins, 2019). Collectively, connected people and devices were generating 3.5 quintillion bytes of data every day in 2023 (Wise, 2023). The volume, variety, and richness of personal organic data are such that they may soon be routinely included alongside self-reports of behavior in psychological research (Montag et al., 2022). A significant advantage of organic data is that, for many sources, it is available across the globe, making possible national and cultural comparisons that would be too costly or logistically complex for researcher-managed data collection. In particular, organic data may prove useful for characterizing environments and situations at a level that could never be achieved using observers and coders (e.g., Suel et al., 2019).

Numerous types of organic data are generated by a growing number of sources that pervade nearly all aspects of everyday life. Virtually all business transactions produce a digital trace; examples include credit card use, utility use, health service access, and product scans. Social media activity is a rich source of information about social experience reflected in photographs, short videos, comments, and reactions to social media posts. Information generated by routine mobile device and/or computer use ranges from sensor-based information, such as location and movement, to features of personal experience, such as search history, information captured by mobile apps, text messages, and photos and videos. Increasingly, information is generated by objects and equipment that collectively constitute the Internet of Things, from which the Internet of Behavior is derived (Sun et al., 2022). Among the "things" connected by embedded technology that transmit information through the Internet are appliances, automobiles, voice assistants, home health monitors, and security systems. In addition to serving the functions for which they were designed, these sources provide rich data on choices and behaviors relevant to constructs of interest to social psychologists.

Examples of organic data use in research by social psychologists suggest its promise as a source of information about constructs across targets. Observable Facebook profile information, such as the number of "friends," the number of posts, and the number of photographs correlates moderately to strongly with self-reports of extraversion (Gosling et al., 2011). A model based on more than two million time stamps on end-of-day Reddit posts produced scores strongly correlated with users' self-reported bedtimes (Meyerson et al., 2023). Perceivers' ratings of 𝕏 (formerly Twitter)-using peers using only a subset of their posts were moderately accurate when compared to self-reports of impulsivity (Orehek et al., 2017). An analysis of the use of Google search terms that indicate interest in high-risk/reward ways of making money (e.g., "lottery") by U.S. states revealed a moderate correlation between online searches and income inequality as reflected in states' Gini coefficient (Payne et al., 2017). Machine learning algorithms applied to digital traces from Facebook activity accurately distinguished between users who tend toward problematic social media use and those who do not (Marengo et al., 2022). As is evident in these examples, the organic data of most interest to social psychologists have been digital traces from social media activity. Moreover, those data have typically been at the individual level and coupled with designed data used to interpret and validate digital phenotypes. Research that uses organic data to characterize populations and features of the environment is rare; however, as training in the use of data management and

analytic tools for large data sets becomes more accessible, uses of organic data that depart from standard practices with designed data will increase.

A significant barrier to using organic data effectively for measuring constructs is the extraction of scores that correspond to constructs of interest. Most forms of organic data are voluminous and not generated to measure constructs. The challenge, then, is how to use organic data to produce construct-relevant scores that are reliable, valid, and reproducible. An initial consideration is whether and how data of interest can be accessed. Digital traces from some sources are publicly available (e.g., Reddit), but access is controlled for many sources (e.g., 𝕏 [formerly Twitter] data are available only with Academic Research access). When accessibility has been arranged, the next challenge is how to move (if allowed) and work with a data set that is several orders of magnitude larger than the typical "large" data set in social psychology. Organic data may require a level of screening, cleaning, and manipulation that exceeds what is typical of designed data in social psychology. Manual procedures are not feasible for this work; instead, automated algorithms are necessary to detect unusable data (e.g., activity attributable to bots; Orabi et al., 2020) and extract meaningful information (Cafarella et al., 2009). Cleaned data are then suitable for applying algorithms and analytic methods for digital phenotyping, descriptive visualizations, and prediction (Yarkoni & Westfall, 2017). A number of validity concerns specific to the use of organic data for measuring constructs must be considered (for a review, see Xu et al., 2020). These range from errors in algorithmic outputs (van Giffen et al., 2022) to decisions about algorithmic parameters and procedures before examining the data (Hofman et al., 2017) to measurement bias attributable to biased algorithm training (Tay et al., 2022). Furthermore, company algorithms that control what content social media users are shown (e.g., Meta Business Help Center, n.d.) or recommendations they are provided (e.g., YouTube, n.d.) may result in usage data that do not reflect personal decisions and potentially distort personal preferences. These validity concerns are specific to the uses of organic data and add to the traditional validity concerns relevant to measurement in social psychology. Yet, given the richness, prevalence, and increasing availability of organic data, and the potential to access features of constructs uniquely available in such data (e.g., exposure to misinformation spread primarily on dark platforms; e.g., Sirola et al., 2022; Zeng & Schäfer, 2021), accessible methods for addressing them are sure to emerge (e.g., van Giffen et al., 2022).

SELF-REPORT MEASUREMENT

Despite a large and growing number of sources of information for measuring constructs, social psychologists rely heavily on self-report measurement (e.g., Boyle et al., 2015; Weidman et al., 2017). Self-reports can be elicited using open-ended or free-response queries; however, far more common are statements or questions followed by discrete response options from which respondents choose one. Given the widespread use of closed-ended items in social psychology (Sassenberg & Ditrich, 2019) and a large number of considerations relevant to their effective use (Stone et al., 2000; cf. Haeffel & Howard, 2010), the focus in this section is self-report measurement of this type. Readers interested in self-report measurement using free-response strategies will find detailed treatments in other sources (e.g., Cathain & Thomas, 2004; Friborg & Rosenvinge, 2013; Grysman, 2015; Haddock & Zanna, 1998).

Preference for the closed-ended self-report method is seldom justified either in terms of its general merits or as the best choice for a particular research question or context. Several features of the method likely have led to its popularity and uncritical use. Foremost is the simplicity of data collection, a feature that has grown in appeal with the increasing availability and sophistication of

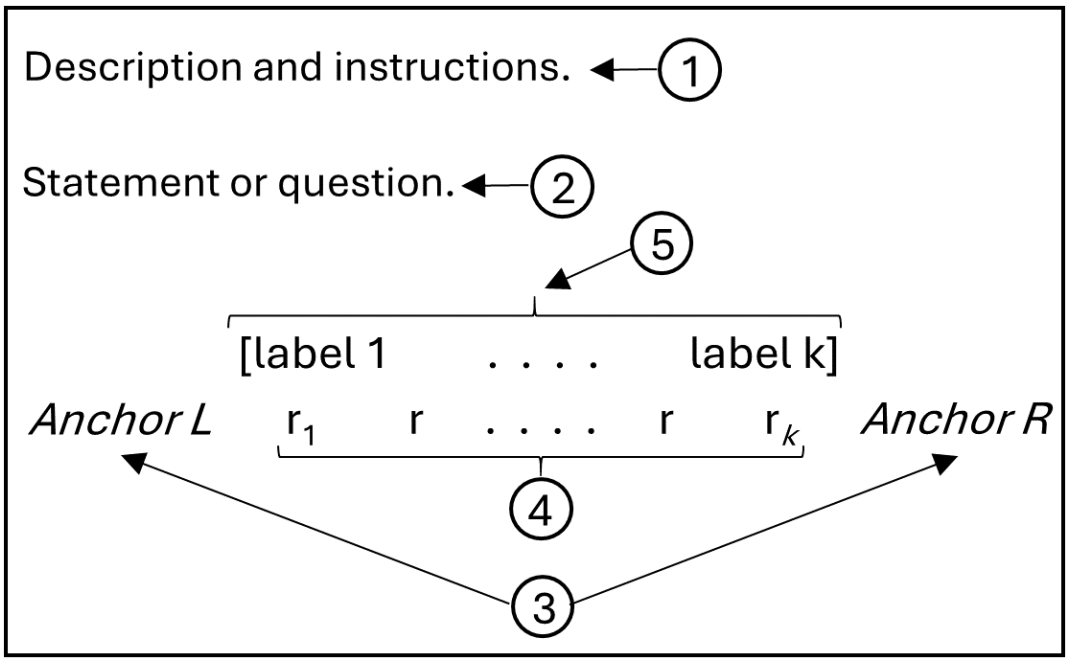
remote methods of self-administration. Items are easily administered to potentially large and distributed samples, requiring little or no interaction between respondents and research personnel. Moreover, with rare exceptions, administration does not require specialized space or skill. Another appealing feature of self-report measurement is that, for many constructs, participants are the most appropriate source of information. For example, constructs such as attitudes, beliefs, and opinions are people's personal subjective views of objects, ideas, and preferences. Finally, self-report measurement is assumed to be straightforward and simple, requiring nothing more than generating a set of direct questions or statements followed by response options. Ready access to web-based tools, such as SurveyMonkey, Google Forms, and Qualtrics has made it possible for someone with little or no knowledge of measurement to develop and administer a self-report questionnaire. However, as the number of topics covered in this section attests, effective self-report measurement requires substantial knowledge and expertise.

The appealing features of the closed-ended self-report method are offset by several limitations and drawbacks. The most significant limitation is that people are unable to validly report their own higher-order cognitive processing (see earlier section on cognitive attributes of constructs). Yet, even when respondents are privy to their own standing on constructs of interest, they may be unwilling to report accurately or unable to do so. Potential sources of influence on responses to self-report items other than the focal construct are well-documented and include social desirability (Paulhus, 1984; Ziegler & Buehner, 2009), failure to recognize reverse-scored items (Weijters et al., 2013), a shift in frame of reference across measurement occasions in studies of change (Howard & Dailey, 1979), and a failure to accurately or completely recall prior behaviors and states (e.g., Horvath, 1982; te Braak et al., 2023). Access and bias concerns aside, the complexity of some constructs, as specified by theory, may be difficult to capture in statements or stimuli typical of self-report measurement (e.g., Cording et al., 2010). Additional concerns related to understanding and interpreting statements and response scales in self-report measures arise for research involving low-literacy respondents (Chachamovich et al., 2009). Acknowledging these concerns and the need to attend to them when considering self-reports as a source of information about constructs (Uher, 2023), the remainder of this section addresses features of self-report items that affect their informativeness.

Anatomy Of A Self-Report Item

The typical form of closed-ended self-report items presented to respondents is shown in Figure 4. The encircled numbers identify five parts of an item and its presentation to be considered in developing, adapting, or evaluating self-report items typical of research in social psychology. These parts are referred to when they are relevant to the strategies and concerns discussed below, keeping in mind that not all parts are present in all instances of self-report measurement.

***Figure 4*: A Generic Self-Report Item Showing the Features of Items as Administered That Warrant Careful Consideration in Measurement Design.**



Self-report items and item sets (i.e., questionnaires or "scales") often are prefaced by text that orients the respondent to the items, response format, and potentially the construct they were designed to measure (#1). In addition to making clear how the response scale is to be used, emphasizing confidentiality, and encouraging honest responding, the text in this part may be used to set the time frame (e.g., past 30 days) or context (e.g., when with peers) the respondent is to reference in reporting. The only part frequently receiving serious consideration in item/questionnaire development or evaluation is the statement or question about which a rating will be given (#2). Statements or questions may refer directly to the construct of interest or indirectly to it when respondents' awareness of what is being measured would bias their reporting about it (Paulhus & Vazire, 2007). As with instructions, statements and questions may include information about the time frame or context, often in the form of repeated item stems for every item in a set. Of critical concern but typically receiving little consideration in the development and evaluation of self-report items are parts 3, 4, and 5. Anchors (#3) determine the range of ratings. Although *strongly disagree* and *strongly agree* are common—often used without thought for their appropriateness—many options are possible that affect how scores are interpreted in important ways. Respondents provide ratings that contribute to scores by selecting a value from a set of options (#4) between pairs of anchors. The options may be expressed in explicit numeric terms (e.g., 1 to 7), a continuum that is underlain by numeric values as in digital sliders or visual analog scales, or sets of verbal expressions of quantities. Finally, anchors may be used as labels for the extreme values, and additional labels may be provided for each interim value (#5; e.g., *strongly disagree*, *disagree*, *neither agree nor disagree*, *agree*, *strongly agree*). This feature is optional and, for reasons discussed below, not generally recommended. Issues include finding appropriate labels for five or more response options, finding labels suitable for certain anchors, and labeling of the middle response option. Each of these parts is relevant for the performance of self-report items and, therefore, warrants careful consideration in their design and evaluation (see Schwarz & Oyserman, 2001, for a review of the cognitive processes involved in self-reporting).

The widespread use of multi-item self-report measures in social psychology owes to the seminal work of Rensis Likert on the measurement of attitudes (Likert, 1932). Likert pioneered the general approach to developing multi-item scales still in use by social and personality psychologists: (1) generate candidate items; (2) select the desired number of items that meet selection criteria; (3)

evaluate the reliability and validity of the resultant scale. The term "Likert scale" refers to multi-item measures produced using Likert's steps and format; not individual items or the set of response options as it is sometimes used (Carifio & Perla, 2007). Contemporary multi-item self-report measures are perhaps best described as "Likert-type" scales, as they resemble measures that reflect Likert's recommendations but differ in particulars (Uebersax, 2006). Such scales reflect Likert's general approach for measuring constructs of many types (i.e., not just attitudes) and advances in questionnaire design, administration, and evaluation (Jebb et al., 2021).

Considerations In The Design And Evaluation Of Self-Report Measures

Other characteristics of sets of items beyond those identified in Figure 4 that warrant careful thought include scale length, item order within a set, the inclusion of reverse-scored items, and how items are presented visually. As noted earlier, researchers typically are concerned most—even exclusively—with the statements or questions. This concern is evident in the presentation of scales in research reports as simply a set of statements or questions, perhaps with the suggestion of a response scale. However, the other parts of self-report items, and how sets of items are organized and presented, influence how the construct is reflected in scores and, therefore, how scores are used in description and hypothesis testing. The remainder of this section focuses specifically on five considerations in the design, evaluation, and use of self-report measures. These considerations, which are summarized in Table 3, are, for the most part, independent of the mode of administration, which is not addressed; informative discussions are available elsewhere (e.g., Bowling, 2005; Schwarz et al., 1991).

***Table 3*: Summary of Considerations in the Design and Analysis of Self-Report Measures**

Scale length	• number of items • content coverage
Statements of questions	• appropriateness for target population • contextual information
Continuum specification	• bipolar versus unipolar • analytic concerns
Response options	• optimal number of options • analytic concerns
Response labels	• labeling anchors • labeling response options

Scale Length

When developing new self-report measures or making decisions about how much time is available for the measurement of a construct in a survey or experimental session, the question of how many items to use is relevant. A practical consideration is the typical time constraint in research settings where the measure is likely to be used. For instance, studies that enroll online samples are often in the range of 15-20 minutes, rarely exceeding 30 minutes, in duration. A reasonable expectation is 60 items in a 10-minute period (CloudResearch, n.d.); thus, the typical battery would accommodate 90-120 construct-relevant items (though respondents can sustain attention in responding to as many as 300 items; Bowling et al., 2022). Constructs for which the only validated measures are lengthy are not suitable for this commonplace research setting. It is perhaps unsurprising that abbreviated versions of lengthier measures of constructs have been developed and, in many cases, have become the most frequently used measure of the construct (e.g., Duckworth & Quinn, 2009; Hoyle et al., 2002; Tangney et al., 2004). This trend toward shorter scales is unlikely to change, suggesting that the target length of new measures should be consistent with widely used abbreviated versions of longer scales.

A fundamental conceptual consideration with implications for scale length is whether an item set fully covers the content of the targeted construct as specified by theory. This determination requires a thorough delineation of the construct against which the content of items in a set can be compared. That comparison could inform the development or use of an item set in several ways. If the conceptualization of the construct specifies several attributes, then one or more items could be developed for each attribute. More than one item per attribute might address different contexts in which the attribute is evident or different forms of expression. For complex constructs, this approach to content coverage can result in a lengthy scale inconsistent with the principle that shorter is better for most research settings. In such cases, getting adequate content coverage with fewer items requires statements or questions that are more general or abstract, referencing clusters of attributes or allocating only one item per attribute. The most extreme form of this move from granular to general content coverage are measures that purport to cover content with one or two items. Examples include one- and two-item measures of the Big Five personality domains (Gosling et al., 2003; Woods & Hampson, 2005), a one-item measure of global self-esteem (Robins et al., 2001), and one-item measures of life satisfaction (Gnambs & Buntins, 2017). The value of such measures for many research settings has increased attention to developing single-item measures (Allen et al., 2022; Fisher et al., 2016). Construct coverage is a particular concern when brief forms of longer measures are produced by eliminating items based on empirical criteria (e.g., Koczkodaj et al., 2017). Unless multiple items per attribute were included in the longer scale, the shorter scale would ignore some attributes of the construct and, therefore, not fully reflect it. In such cases, a better approach would be the development of a smaller set of items that reflect the construct in more general terms. In any case, very brief measures should be compared to longer and more complete measures of the same construct to ensure equivalent measurement of the construct (e.g., Gnambs & Buntins, 2017).

Statements Or Questions

Regardless of the number of items in a set, the characteristics of the statements or questions to be rated are a key consideration. Good treatments of structural features (e.g., statement/question length) and specific wording concerns (e.g., reading level, use of negations) are available in the literature (e.g., Blanton & Jaccard, 2018; Boeteng et al., 2018). Here, the focus is on more general

considerations, assuming care has been taken to optimize the structure and wording of statements or questions to be rated.

The most fundamental concern is the degree to which the statements or questions reflect the target construct as they are read and understood *by respondents from populations for which the scale will be used*. A statement that accurately reflects a construct to knowledgeable researchers may not invoke the target construct or may invoke other constructs when read and rated by respondents. For example, statements that refer to being anxious may suggest eagerness to some respondents and intense worry to others. Two scale development strategies mitigate against this threat to validity by involving typical respondents in the statement/question design process. An efficient means of ensuring that candidate items are read and understood as desired is to engage members of the target population at the stage of item generation. *Qualitative interviews* and *focus groups* are means by which representatives of the target population can convey their understanding of the target construct, specific ways the construct manifests, and ways of expressing differences between people on the construct (e.g., Nassar-McMillan et al., 2010). Engaging representative population members in this way may suggest statements, descriptions, and themes that can be used to generate candidate items. Alternatively, researchers can generate candidate items based on a well-developed conceptualization of the construct before inviting input from members of the target population. **Cognitive interviewing** may be used to elicit input aimed at ensuring statements or questions are clear, understandable, and accurate in their representation of the construct for the population (Beatty & Willis, 2007). Whether elicited before or after item generation, input from the target population can contribute to better self-report measurement by increasing the probability that the respondents will reference the target construct when choosing responses.

When considering a statement or question before choosing a response, respondents are likely to consider contextual information related to time and place or specific attributes of the construct. They might also bring to mind other people or groups, considering their own standing on the construct by comparison. Collectively, these are matters of *reference*. When references are unspecified, the references respondents freely choose may vary widely, producing variance not attributable to the target construct (e.g., Lira et al., 2022). The most commonly provided reference is the time period over which the respondent is to report. When the period is deemed irrelevant, items might begin with "Generally" or "Typically," presumably eliciting information about respondents' standing on a form of the construct that is temporally and cross-situationally consistent. Alternatively, because of study design, as in longitudinal studies, or the nature of certain constructs, specific information about the time period to be referenced might be specified. Examples are "In the past month," "Since you last reported," and "Right now, at this moment." Such references influence retrieval strategies for personal behaviors or experiences relevant to standing on the construct (Walentynowicz et al., 2018). Other frames of reference include context (e.g., home, work, online) and comparison others (e.g., peers, significant others). When manifestations of the construct that differ across contexts or comparison others is of substantive interest, a useful strategy is to include a word or phrase in the statements or questions that refer explicitly to the reference and can be tailored. For example, an item on the Perceived Cohesion Scale (Bollen & Hoyle, 1990) reads, "I see myself as part of the _____ community." The blank could be replaced by the name of a school, a town, or a workplace; the selected name would be included in all items. Items of this sort assume that standing on the target construct is tied to a specific context, to which respondents should be referring when self-reporting.

A final consideration is whether any of a set of items should be phrased to reflect the absence or opposite of the target construct. Such items might take two forms. A common approach is to insert a

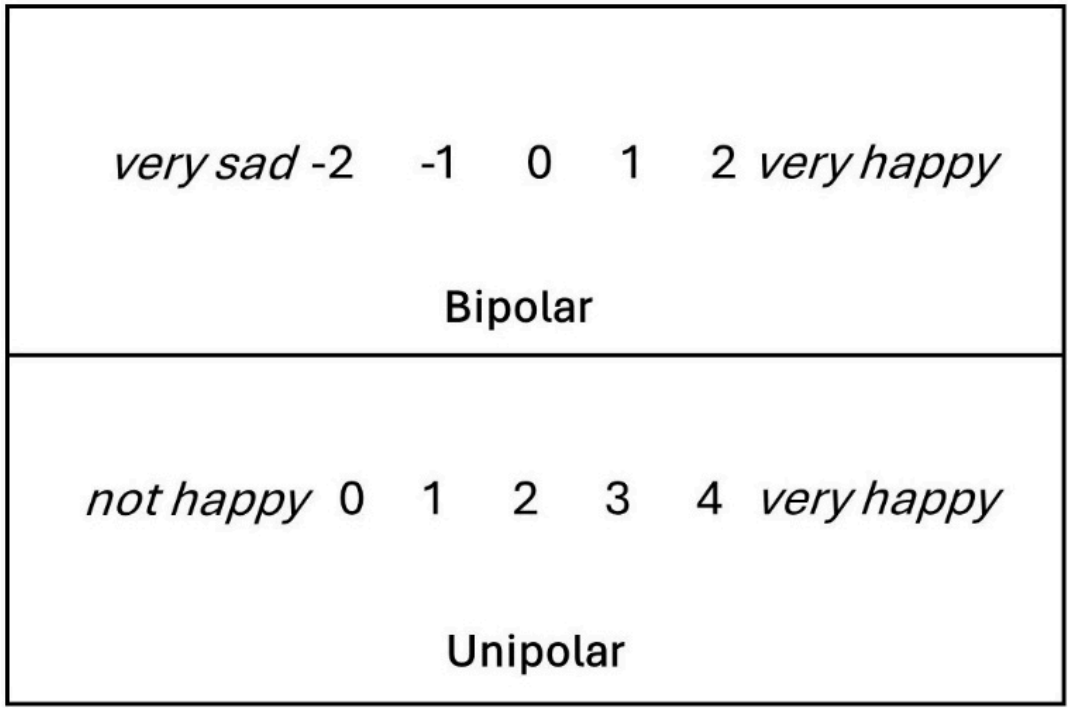
negation such as “not” or “don’t” into a statement that reflects the construct. Lower scores on such items—that is, not endorsing the absence or opposite of the construct— are assumed to reflect the construct. An alternative approach is to state or inquire about an attribute that is the opposite of the attribute of interest, in which case lower scores indicate that the opposing attribute does not apply. A challenge in writing these items is a clear understanding of the opposite or “low pole” of constructs and their attributes. The risk is stimulating responses that refer to aspects of the opposing construct that are not, strictly speaking, relevant to the construct of interest. Findings from comparisons of responses to items with and without negation or reference to opposing constructs indicate that reversed items do not alleviate the biases they are designed to address, instead creating confusion. Fewer mistakes due to inattention and confusion are evident when all items are worded in the direction of the target construct (van Sonderen et al., 2013), and method effects attributable to wording direction are less likely to emerge in psychometric evaluations (Zeng et al., 2020). In short, reversed items should be used sparingly, if at all, with a full understanding of and accounting for their influence on scores (for a review and integrative model, see Weijters & Baumgartner, 2012; Weijters et al., 2013).

Continuum Specification

The assignment of scores to people concerning a target construct assumes that the construct has a theoretical continuum. Consideration of how item responses array people along the continuum is known as *continuum specification*. It describes the process through which the continuum is defined and operationalized in self-report measurement (see Tay & Jebb, 2018). The two main considerations in continuum specification are (a) defining the target construct continuum, and (b) operationalizing the continuum based on the proposed definition.

One aspect of defining the construct continuum is to consider the polarity of a construct—in most cases, whether the construct is bipolar or unipolar. As shown in Figure 5, a *bipolar continuum* has anchors that convey opposing attributes or states. For example, affect valence construed as happiness-to-sadness is a bipolar continuum. In contrast, a *unipolar continuum* has the lower pole representing the absence of an attribute or state rather than its opposite. For instance, positive affect construed as the presence of happiness to the absence of happiness is a unipolar continuum (see Russell & Carroll, 1999). How the polarity of constructs is defined, and whether the response format matches those definitions, is an often-overlooked aspect of measurement validity.

***Figure 5*: Bipolar and Unipolar Continuum Specifications for Measuring Happiness**



Well-defined polarity of constructs allows for the rigorous operationalization of the construct continuum. For one, specifying the construct continuum affects whether reverse-worded items should be considered (concerns raised earlier, notwithstanding). When the construct continuum is unipolar, including reverse-worded items does not align with the content of the construct. For instance, it would be inappropriate to include a statement or question that refers to being very sad in a measure of the unipolar positive affect construct, in which the lowest pole represents less happiness rather than being very sad (Russell & Carroll, 1999). Another aspect of operationalizing the construct continuum is the consideration of response scale anchors (#3 in Figure 4). For example, some response scales explicitly assume a bipolar continuum by including an anchor that corresponds to the opposite of the construct (e.g., top panel in Figure 5). Such response scales do not align with a unipolar continuum. For unipolar continua, anchors should be used that indicate the presence or absence of the construct (e.g., bottom panel in Figure 5).

The consideration of construct continua is not merely one of preference. A failure to consider this feature of self-report items can create significant inferential problems and confusion. For instance, using unipolar scales results in a theoretical maximum correlation between happiness and sadness of -0.47 rather than -1.00 (Russell & Carroll, 1999). Moreover, evaluations of latent structure can indicate two factors, one of which is a statistical artifact, contributing to the longstanding controversy of whether positive and negative emotions are distinct or part of a bipolar continuum (Tay & Kuykendall, 2017). This concern may apply to other areas of research, such as attitudes, personality, and interests (Tay & Drasgow, 2012). Moreover, formalized theoretical models of the interaction between attributes of a construct behave very differently for bipolar and unipolar continua (Haslbeck et al., 2021). More generally, the psychometric properties of unipolar and bipolar scales differ enough that they should not be viewed as interchangeable (Höhne et al., 2021, 2022).

Number Of Response Options

Although related to continuum specification, the consideration of how many response options to offer respondents (#4 in Figure 4) warrants separate attention, given its importance for scoring, interpretation, and analysis. The number of response options is often given little thought in social-

psychological research, governed more by precedent and norms than careful consideration. The considerations that influence how many options to offer respondents are of two general types: (a) implications for how respondents understand and locate themselves on the response continuum and (b) implications for scoring, modeling, and analysis. The decision about how many response options to offer and how to present them should balance these two considerations.

The most basic concern in choosing the number of response options is ensuring that people who differ in their standing on the construct differ in their self-rating. Related to this concern is ensuring that different ratings are attributable to differences in standing on the construct and not random error. The possibilities for the number of response options range from coarse (*yes-no*) to practically continuous (e.g., digital slider). From a respondent understanding and response selection perspective, a rationale for choosing a number is the minimum number of options that allows respondents to unambiguously express their standing on the construct without requiring them to choose between options they cannot differentiate on the construct. The ideal balance results in variance attributable primarily to the construct as expressed in the statement or question, with little or no error variance attributable to an inability to differentiate between adjacent values. Although more response options might produce more variance (a common justification), that variance may not be attributable to the construct. As an example, respondents given response options ranging from 1 to 15 might find it difficult to choose between intermediate values—say, 4, 5, and 6—because it is unclear how they differ in reflecting standing on the construct. An additional consideration of this type is whether to offer an odd or even number of options, the former providing a midpoint. If response options are not labeled, respondents may view the midpoint as just another choice along a continuum. If options are labeled, then how midpoints should be labeled is a concern. For instance, when measuring an attitude, the midpoint could be labeled *neither agree nor disagree*—an expression of indifference—or labeled *agree and disagree*—an expression of ambivalence (e.g., Kaplan, 1972; Yoo, 2010). As part of the continuum specification process, the choice about whether to offer a midpoint and, if labels are provided, how to label it, should follow from the conceptualization of the target construct.

A second consideration is statistical and concerns how the results of analyses using scores are affected by the number of response options. If underlying a response scale with a finite number of options is a continuous scale, then one concern is how coarsely categorizing the scale affects associations estimated using scores. It can be shown through simulation (e.g., Bollen & Barb, 1981) or comparisons of associations for respondents given different numbers of response options (Simms et al., 2019) that reducing a continuous scale to a relatively small number of points along a continuum —**coarse categorization**—attenuates statistical associations with other variables. For example, a population correlation of .20 estimated from coarsely categorized scales will be .12 with two categories (e.g., *yes*, *no*) and .18 with five (e.g., *strongly disagree* to *strongly agree*). The attenuation of estimates of association is increased further when the intervals between categories are unequal (Onoshima et al., 2019). Assuming, however, that, for reasons outlined in the previous paragraph, coarse categorization is warranted, a question that arises is whether scores can be treated as continuous in analytic models that assume continuous measurement scales. When normality cannot be assumed, simulations suggest that five or more categories are sufficient to justify using estimators that assume scores were generated by continuous scales (Rhemtulla et al., 2012; Robitzsch, 2020).

Response Labels

Related to the number of response options are considerations about the use of labels in combination with, or instead of, numeric options (Maitland, 2009). Referring back to Figure 4, labels might be associated with response options in two ways. First, and most common, labels anchor numeric scales, sitting alongside the most extreme numeric options (#3). Second, labels for the most extreme responses are associated with those numeric options, and additional labels are associated with interim numeric options (#4). A version of this option, made popular by Web-based survey tools, dispenses with the numeric options altogether, presenting the respondent with only the labels. Regarding anchors, the choice of labels should consider how statements or questions are phrased and how the construct is expressed. *Strongly disagree* and *strongly agree* are common choices; however, they make assumptions about the underlying continuum that are rarely examined, as well as the wording of statements. Other anchor options define the continuum as one of degree (*not at all–very much*) or frequency (*hardly ever–nearly always*) of expressing the construct. When the anchors refer explicitly to the construct, the extremity is a consideration. For example, anchoring a scale with *extremely sad* and *extremely happy* may discourage the choice of the lowest and highest values, given the unlikelihood of extreme expressions of these emotions. This concern must be balanced against the need to differentiate respondents at the highest and lowest levels of the construct from respondents at the middle of the range.

Rather than providing only anchors and no labels on the response options themselves, the scale may provide a label for each response option. A widely used form of this labeling is *strongly disagree*, *disagree*, *neither agree nor disagree*, *agree*, *strongly agree*. Although the concern about the labeling of the midpoint raised earlier applies, labels in this range do not suffer from an issue for other ranges—how to communicate degree using labels that define equivalent intervals. The issue is particularly challenging with more than four or five response options. Nevertheless, research on labeling versus not labeling response options suggests that using labels for measuring some constructs in some populations is preferable (for a review, see Krosnick & Fabrigar, 1997). Beyond the simple labels typical of survey items, an expanded labeling format that embeds the typical labels in different forms of the statements to be rated (e.g., "I am very lazy," "I am somewhat lazy," "I am not at all lazy") has shown promise for use with some measures (Zhang et al., 2023). With more than a few response options or a continuum for which interim labels are not easily generated, numeric scales with anchors are the better option and do not appear to produce lower-quality data or compromise validity when compared to the labeled response format (e.g., Lewis, 2019).

Self-Reports In Intensive Longitudinal Studies

To this point, the focus has been on self-report measurement in the context of cross-sectional survey or questionnaire studies. Another research context in which self-report measurement is routinely used is *intensive longitudinal studies*, in which brief questionnaires are administered repeatedly—often multiple times per day—over a short period of time (e.g., two weeks). This experience sampling methodology (ESM) has as its goal capturing self-reported momentary experiences using diary methods (Bolger et al., 2003) and ecological momentary assessment (Shiffman et al., 2008). ESM can improve self-report measurement of constructs in several significant ways. First, it enables the assessment of within-person phenomena on which substantial psychological theory and interventions are based (e.g., enacting behavior X leads to behavior Y). Second, it enables more accurate measurements by overcoming memory biases in retrospective or global self-reports (Scollon et al., 2003). Third, the dynamic nature of constructs can be measured, enabling the capture of natural-occurring variability in constructs such as affect (Kuppens et al., 2010), personality (Fleeson & Gallagher, 2009), and performance (Dalal et al., 2020). Although most

of the considerations relevant to survey or questionnaire studies apply in the ESM context, limitations of time, attention, and presentation format give rise to additional considerations.

A key consideration in the use of ESM is the time required to complete each questionnaire. Although for an ecological momentary assessment study, the total time required to respond to as many as five requests per day over the course of 14 or more days is not substantial, repeated requests are an intrusion increasingly likely to produce reactance and nonresponse as the magnitude of the intrusion increases. It is recommended to keep the questionnaire duration short (5-10 minutes), increasing brevity (perhaps down to 1-2 minutes) as frequency increases, for a total response burden during the day of 15-20 minutes in total (Fisher & To, 2012; Tay, 2018). Increasing frequency rather than questionnaire length is preferable for reducing participant burden (Eisele et al., 2022). Finally, regarding study duration, many ESM studies typically last one to two weeks, although some span months (Christensen et al., 2003; Fleeson & Gallagher, 2009); study duration contributes to participant burden and fatigue. Significant concerns also arise from changes to responding due to familiarity (e.g., response sets bias), which may be ameliorated by randomizing questionnaire items (Swendsen et al., 2000). The measurement challenge is how to adequately assess the constructs of interest quickly. Assuming the estimate of 10 secs per item, the limit may be as few as 15-20 items for all constructs of interest (plus additional measurement through passive sensing; Thapa et al., 2021). As a result, constructs are frequently assessed using single items (e.g., Cloos et al., 2023; Song et al., 2023). Single-item measures are available for a growing number of constructs (Allen et al., 2022) and provide an efficient and effective means of measuring constructs in the intensive longitudinal research context, where the focus is the state form of the construct.

Although the timing of administration is relevant for traditional surveys, especially in longitudinal studies, it is of central importance in intensive longitudinal studies. Not only is timing relevant to concerns related to intrusiveness and burden, but it also applies to how scores on measures are used to operationally define constructs and their expression. Thus, when using self-report measures with ESM to capture the dynamic, temporal, and contextual aspects of constructs, considerable thought must be given to the rationale for when and how often to request responses from participants. Four types of schedules are typical when using ESM, each with strengths and weaknesses (see reviews by Christensen et al., 2003; Fisher & To, 2012). A **time-contingent (or interval-contingent) schedule** requires participants to complete measures at fixed times each day. It is applied when there are well-established rhythms or timings to behavior changes expected in the sample of interest. A **signal-contingent schedule** specifies the random timing of signals during the day. Typically, the randomization occurs within time slots to intersperse self-reporting. This schedule can increase the validity of assessing broad life experiences as classically used in ESM (Csikszentmihalyi et al., 1977) and is also used to capture fluctuations that may unfold idiographically. An **event-contingent schedule** relies on participants to complete self-report items when a specific event occurs. This schedule is helpful for assessing episodic behaviors that arise in events, such as social interactions or substance use (see review by Ebner-Priemer et al., 2009). Finally, there is a **context-contingent schedule**, wherein notifications are sent to participants based on passively sensed information of surrounding context (e.g., proximity to home or office; Das Swain et al., 2019). Because sensors consistently detect contexts, they reduce measurement gaps characteristic of event-contingent schedules that rely on participants to remember to report an occurrence.

Taken together, the timing and frequency of measurement in intensive longitudinal studies should allow for valid assessment of the dynamics of the focal constructs with sufficient data points to reliably estimate effects at specific levels of analysis (e.g., day-level changes, week-level changes). Moreover, more complex calculations of temporal constructs and their interrelations—variability,

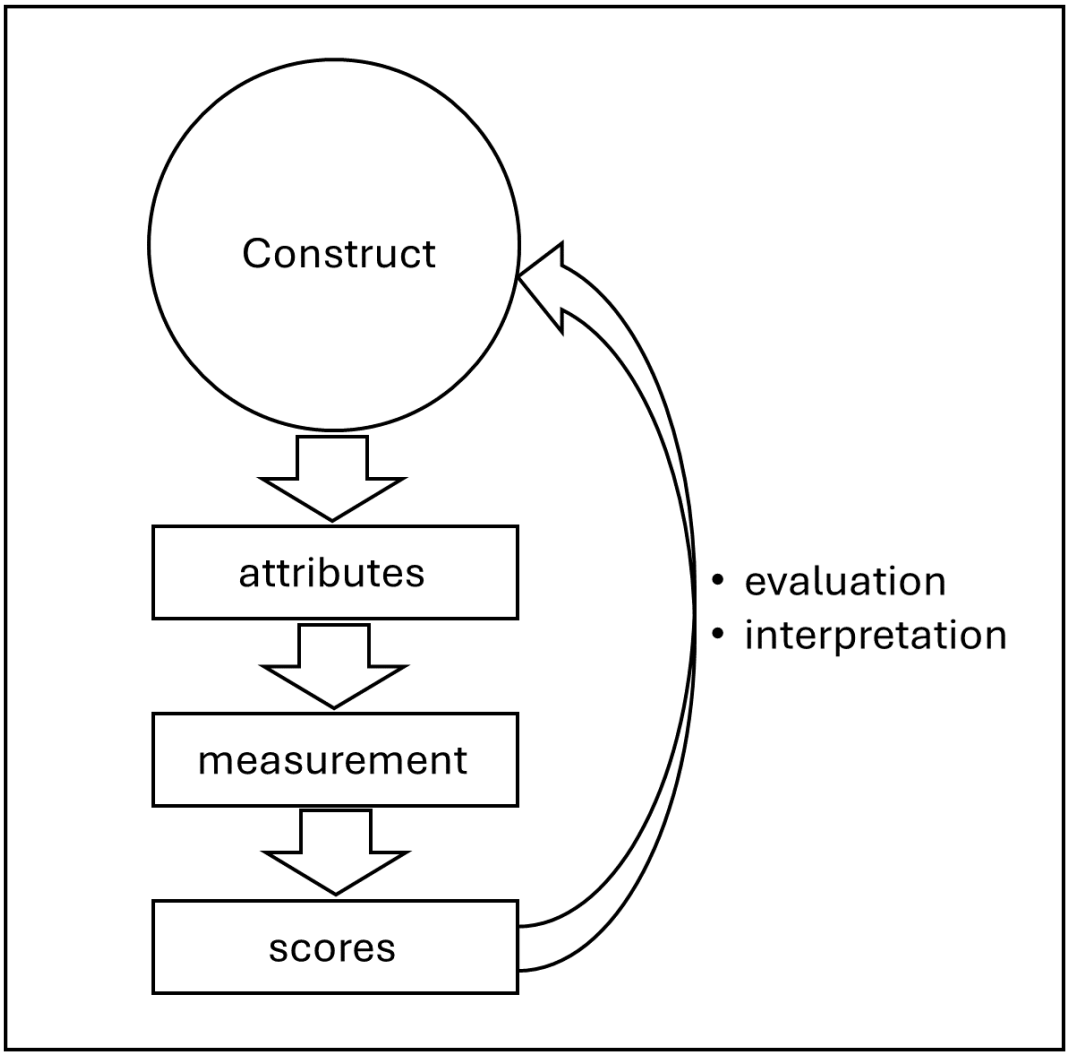
reactivity, inertia, cycles, and feedback loops—require sufficient and appropriately timed measurements to calculate them validly and reliably (Diener et al., 2020). Careful consideration of the specific types of construct-relevant parameters implicated in the motivating questions or hypotheses is necessary for the right number of assessments and their frequency.

PROPERTIES OF SCORES

Scores are the product of measurement in quantitative research. Reflecting the widespread use of self-report measures, the preponderance of information about working with scores in social psychology refers to scores derived from self-report measures, frequently referred to as "test scores" (e.g., Kane, 2013; Messick, 1989). Because self-report measurement is but one of many potential sources of information about constructs, and constructs may refer to targets other than people, language that does not assume scores in a set used to create a composite score are responses to self-report items is used in this section. The origin of scores may be any of the information sources reviewed earlier, including specialized equipment, digital devices, and the many forms of organic data. Moreover, a composite may comprise scores from multiple sources, deviating from the general bias toward monomethod measurement. Thus, *scores* refers to numeric values produced by or extracted from any of the wide array of sources of information reviewed earlier.

The evaluation and interpretation of scores are relevant for judging the quality and usefulness of the source from which they derive; however, the source and the scores derived from it are not to be treated as one and the same (Kane, 2013). Any interpretations or inferences offered concerning a measured construct apply specifically to a set of scores from a particular sample at a particular time and place. This relation between scores, measurement, and constructs is depicted in Figure 6. Scores are evaluated and interpreted, and inferences are drawn about their associations with other scores, with reference to a targeted construct. Whether those evaluations, interpretations, and inferences would apply to a set of scores from a different sample, or even from the same sample at a different time or place, is uncertain. Thus, assertions about reliability and validity when measuring constructs are assertions about those properties as evidenced in a specific set of scores acquired using a specific procedure in a specific research design context. For that reason, generalizability is always a consideration when evaluating and interpreting scores.

Figure 6. **Scores Are Produced by Measurement of Attributes Specified by a Construct, and the Evaluation and Interpretation of Scores Refer Back to the Construct.**



Reliability

A fundamental property of scores derived from a measure is their reliability. A *reliable measure* consistently produces scores on the construct it measures. Consistency may be evaluated with respect to the time or place of measurement (e.g., test-retest reliability), items in multi-item scales (e.g., internal consistency), or ratings provided by judges or informants (e.g., inter-rater agreement). Reliability is necessary but not sufficient for valid measurement; scores derived from a measure can reliably measure constructs that differ from the construct of interest. Strategies that aim to maximize certain forms of reliability of scores derived from multi-item scales (e.g., coefficient alpha) may work against valid measurements of the targeted construct (Clifton, 2020). As such, measures to achieve consistency of scores must take care not to undermine the validity of scores as judged with reference to the targeted construct. When valid measurement has been established, higher reliability of scores is important for accurate estimates of population effect sizes (Trafimow, 2016) and fully controlling for confounding constructs when estimating partialed effects (Westfall & Yarkoni, 2016).

Estimates of reliability provided in social psychology research reports are typically values of Cronbach's (1951) *coefficient alpha*, the limitations of which are well-documented (e.g., Cronbach & Shavelson, 2004; Green et al., 1977; Green & Yang, 2009; Schmitt, 1996; Sijtsma, 2009; cf. Raykov & Marcoulides, 2019). Those limitations stem from a set of assumptions rarely met in practice. Foremost among those is the fact that, for alpha to be a point estimate of reliability, one requires the

assumption of *tau equivalence*: the true item scores (which may or may not reflect the targeted construct) are linear translations of each other (Lord & Novick, 1968). Coupled with the assumption of unidimensionality, tau equivalence is evidenced by equal factor loadings on a single factor. Additional commonly made assumptions are uncorrelated errors and normally distributed scores. Failure to meet these assumptions can result in distorted estimates of reliability (Sheng & Sheng, 2012; Sijtsma, & Pfadt, 2021). A more fundamental assumption when interpreting internal consistency as a point estimate of reliability is that the sources of the scores contributing to a composite are correlated and that correlation is attributable to a single latent variable. Referring back to Figure 3, that assumption is reflected in the conceptualization depicted in Panel A. Conceptualizations depicted in Panels B and C make no such assumption and, as such, consistency as assessed by alpha and other indices based on a common factor assumption may not be a relevant form of reliability for these cases (for a discussion of the broader potential impact of this assumption when it does not hold, see Rhemtulla et al., 2020).

It is important to note that, even when the strict assumption of tau equivalence is not met, alpha can still be useful, if it is interpreted as an estimate of a lower bound to reliability. This interpretation does not require the above assumptions and holds for any composite that can be decomposed into a random (measurement error) and systematic (true score) component; whether these components are consistent with a unidimensional model is irrelevant to the proofs that establish alpha as a lower bound (Sijtsma & van der Ark, 2020). When unidimensionality cannot be assumed, reliability cannot be interpreted as precisely as for unidimensional constructs, but may still be profitably conceptualized as the squared correlation between the observed total test score and the true total test score, i.e., the expectation of the total test score (Guttman, 1945; Sijtsma, 2009; Sijtsma & Pfadt, 2021). Cronbach's alpha establishes a lower bound to that quantity; if for a set of items Cronbach's alpha equals .70 in a population, then the squared correlation between the observed total test score and the true total test score is at least .70. This lower bound property is true for any total test score, regardless of whether the composite is unidimensional. As the relations between items approach tau equivalence, the lower bound will approach the actual reliability more and more, and when the relations between items are tau-equivalent, alpha is no longer an estimate of a lower bound to reliability but a point estimate.

Alternatives to coefficient alpha include other internal consistency estimates and other approaches to estimating reliability when internal consistency is not meaningful (e.g., single-item measures, formative indicators). The internal consistency alternatives, which assume a common factor model but drop restrictive assumptions such as equal factor loadings, include the composite reliability for congeneric measures model (Raykov, 1997), variants of coefficient omega (Flora, 2020; McDonald, 1999; cf. Cho, 2021), and the greatest lower bound (Sijtsma, 2009; Woodhouse & Jackson, 1977; for comparisons, see Cho, 2016; Revelle & Zinbarg, 2009; Trizano-Hermosilla & Alvarado, 2016). Omega coefficients perform particularly well with measures typical of social-psychological research (Dunn et al., 2014; Revelle & Zinbarg, 2009; Trizano-Hermosilla & Alvarado, 2016). When, for composites, the common factor does not apply, and for single-indicator measures such as those used in ESM studies or extracted from passive sensing or organic data, consistency typically is evaluated over time by correlating scores from a single sample on multiple occasions (e.g., Fisher et al., 2016; Polit, 2014; Spörrle & Bekk, 2014). Although Pearson's *r* is typical for estimating test-retest reliability, the intraclass correlation coefficient is often a better choice (Shrout & Fleiss, 1979; Weir, 2006). The selection of time intervals between assessments is critical and should be informed by an understanding of whether levels reflected in scores on the targeted construct would be expected to change or remain the same at different intervals.

Validity

Validity is a core topic in measurement theory and evaluation (Cronbach & Meehl, 1955; Kane, 2006; Messick, 1989). However, in contrast to properties of scores such as reliability and measurement equivalence, the conceptualization of validity has not converged on a commonly accepted definition over the past century (Newton & Shaw, 2014). At least four conceptualizations, summarized in Table 4, are in use.

***Table 4*: Conceptualizations of Validity**

Properties of the measurement approach	• Does the measurement approach produce scores that reflect the intended construct?
Properties of scores	• Do scores correlate (or not) as expected with criteria?
Interpretations of scores	• Is an interpretation of scores justified given relevant theoretical propositions?
Uses of scores	• Can uses of scores for action and decision-making be justified on political or ethical grounds?

Properties Of The Measurement Approach

Validity can be viewed as a question about the measure used to derive scores: Does it measure what it purports to measure (Kelley, 1927)? For instance, do scores on the Implicit Association Test (IAT) measure individual differences in implicit cognition (Greenwald et al., 1998)? Validity of measures typically involves the causal question of whether the targeted attribute (e.g., an implicit stereotype) causes systematic differences in scores derived from the measure (IAT scores; Borsboom et al., 2004). As such, a primary locus of evidence for the validity of measures lies in the explication of processes that contribute to differences in scores derived from the measurement approach. For instance, what response processes lead individual differences in implicit stereotypes to produce individual differences in IAT scores (De Houwer et al., 2009)?

Properties Of Scores

Consistent with the view that the performance of measures is judged by the scores derived from them (Kane, 2013), validity can refer to the properties of scores. Here, the most relevant property is the correlation between scores and a (possibly latent) criterion of interest (Lord & Novick, 1968). For instance, do IAT scores predict scores on explicit measures of stereotypes or behavior outside the measurement context (Nosek & Greenwald, 2009)? Under the assumptions of classical test theory (Lord & Novick, 1968), predictive validity is limited by score reliability (predictive validity cannot exceed the square root of reliability; Lord & Novick, 1968), and in this sense, reliability implies an upper limit for predictive validity coefficients. However, it should be noted that this property does not generalize to composites (e.g., Panel B in Figure 3). For example, when predicting a criterion from a number of variables, the prediction will be optimal if the predictors are mutually

uncorrelated but highly correlated with the criterion; a composite constructed from such variables (e.g., a total score) will have excellent predictive utility but will not for instance adhere to unidimensionality. Thus, when, as is typical in the field, composing a set of items according to criteria such as internal consistency and unidimensionality, it should not be assumed that the composite is optimizing predictive utility.

Interpretations Of Scores

Validity can be used to refer to interpretations of scores (Kane, 1992; Messick, 1989). Here, the question is to what extent an interpretation of a score is justified. The justification of score interpretations commonly involves the evaluation of a broader theoretical context than merely the psychometric aspects of the situation. For instance, the justification of an interpretation of IAT scores as indicators of, say, latent bias would not just involve a psychometric analysis, but also a definition of bias and a theory of how bias and implicit stereotypes are connected; as such, the validity of score interpretations supersedes (and often encapsulates) that of measures and the scores derived from them (Messick, 1989).

Property Of Score Use

Validity can be viewed as a property of score use, for instance, in educational assessment, psychiatric diagnosis, and personnel selection (Shepard, 1997). Because scores may be used to justify actions, not just interpret them, the validation of score use is not merely a matter of scientific investigation; it also involves the consideration of norms, values, and regulative ideals that guide behavior (Messick, 1989; Shepard, 1997; Wijsen et al., 2022). For instance, whether the use of IAT scores in bias-reduction efforts can be justified depends not only on scientific considerations but also on political ideals and social values.

Although various attempts at integrating or aligning the different meanings of validity have been made (Kane, 2006; Markus & Borsboom, 2013; Messick, 1989), the field does not appear to be converging on a single conceptualization (Newton & Shaw, 2014). As such, validity is a multifaceted property of measurement approaches, and the scores they produce can have different relevance or meaning to different researchers in different contexts. Across all measurement approaches and scores, the primary consideration is whether, as shown in Figure 6, scores can be interpreted or applied with reference to the construct of interest. Because delineations of constructs evolve and the use of measures often falls outside the uses considered when initial evidence of validity was produced, validation is best viewed as an ongoing process of ensuring acceptable correspondence between constructs and measures of them (Flake et al., 2017).

Equivalence

The evaluation of *measurement equivalence* asks whether the properties and meaning of scores derived from a measure are the same for different groups or at different points in time. Referring back to Figure 6 and the arrow from scores to construct, the equivalence question is whether the construct is similarly reflected in scores across the range of groups and time points to be studied. If not, between-group or longitudinal comparisons on the construct using the scores are not meaningful. Historically, measurement equivalence received scant attention from social scientists, reflecting an assumption that observed differences or changes in scores indicated true differences

or changes in the focal construct (Davidov et al., 2014). With increasing attention to diversity in samples and interest in cultural differences, the assumption that scores represent the same construct for different groups must routinely be evaluated (e.g., Eid & Diener, 2009; Han et al., 2019). Measurement equivalence is an important condition, as it facilitates meaningful between-group comparisons, especially when the groups to be compared have different histories, speak different languages or dialects, or are exposed to different cultural norms and expectations.

Measurement equivalence can be examined at two levels: psychometric and conceptual. The preponderance of published commentary and research on equivalence refers to evaluations of **psychometric equivalence**, commonly referred to as measurement invariance testing (for a review, see Leitgöb et al., 2023). Invariance testing involves comparing parameters in measurement models of scores that contribute to composites or serve as single indicators of a construct. In the case of measures based on multiple scores, invariance testing moves through a hierarchical set of comparisons, beginning with whether the number of factors and pattern of loadings on those factors are the same between groups and, if the basic form is equivalent, comparing different sets of parameters (e.g., loadings, errors). Strict invariance (i.e., all parameters are equivalent for the groups) is not necessary for informative between-group comparisons of scores (Byrne et al., 1989; c.f., De Beuckelaer & Swinnen, 2011); however, there is general agreement that equivalence of factor loadings and the intercepts of measurement equations is ideal for unambiguous between-group comparisons of composite scores (Davidov et al., 2014; Tay et al., 2015). This is particularly important if researchers interpret observed mean differences between groups in terms of mean differences in constructs (Borsboom, 2006).

An alternative approach is to focus on the construct rather than the psychometric properties of scores. Even strict psychometric equivalence does not guarantee **conceptual equivalence** (Boehnke, 2022), because the fact that the measurement model behaves the same way in different groups does not imply that one is measuring the same construct in these groups. For instance, it is possible that tests of the psychometric equivalence of a set of personality items administered to one group and a set of intelligence items administered to another group, satisfy measurement invariance criteria; all that is required for measurement invariance to hold is that the structure and parameters of the measurement model are the same. Therefore, conceptual equivalence should not be ignored, particularly when comparisons are cross-cultural or when verbal measures are translated for use in different languages. Retaining the semantic content of items across translations, as would be the goal of translation accuracy verified through back translation (Brislin, 1980), does not guarantee an equivalent representation of the construct. Newer translation approaches include focusing on the target construct in the translation process (e.g., Clifton et al., 2023; Harkness, 2007). A different but promising approach for developing new measures to be used in different languages or for different cultural groups is the development of emically informed items for each group designed to achieve equivalence in capturing the focal construct without the restriction of semantic sameness (e.g., Boehnke et al., 2014). Conceptual equivalence also applies to nonverbal measures, such as digital phenotypes and behaviors gleaned from organic data. Scores derived from these sources may also have different meanings for different groups. Focusing on conceptual rather than psychometric equivalence shifts the generalizability concern from the generality of findings to the generality of theoretical conclusions (Mook, 1983), which typically is of greater interest to social psychologists.

Fairness And Bias

Unfairness and bias in measurement can lead to problematic inferences and conclusions. Moreover, the measurement of social-psychological constructs has significant implications because scores assigned to people on dimensions, such as ability, personality, and attitudes are often used to select, promote, or train them for educational and organizational opportunities (Kubiszyn & Borich, 2016; Sackett & Lievens, 2008). For this reason, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) and the *Principles for the Validation and Use of Personnel Selection Procedures* (2018) have been developed to address the key issues surrounding fairness and bias in measurement.

Foremost, fairness is recognized as a social concept with multiple valid perspectives. One perspective stems from a distributive justice standpoint of ensuring equal group outcomes; that is, the same scores are obtained between groups on the same measure. Another perspective stems from a procedural justice standpoint. Fairness concerns of this type include equitable treatment (i.e., people are treated in the same manner in the measurement protocol irrespective of group membership); comparable access (i.e., people are not systematically restricted in their access to the construct being measured); and the lack of measurement and predictive bias. Critically, the *Standards* and *Principles* reject the equal group outcomes definition of fairness for psychological measurement, because, though measures that reveal group differences should be scrutinized, score differences between groups do not necessarily indicate unfairness in measurement. For example, finding differences in scores on a psychological attribute (e.g., well-being, attitudes, mindsets, personality) between groups may reflect systemic differences in socioeconomic status (Brown-Iannuzzi et al., 2017), culture (Triandis & Suh, 2002), and other factors.

Although unequal group outcomes are not necessarily indicative of unfairness in a measurement procedure, the consequences of scores derived from measures that show large group differences, especially for high-stakes decisions, can perpetuate a lack of representation of minority groups. Indeed, there are concerns that measurement procedures are typically validated using WEIRD (Western, Educated, Industrialized, Rich, and Democratic) samples (Henrich et al., 2010); a failure to consider generalizability to other groups can be a form of unfairness (Greenfield, 1997). Scores can also be unfairly interpreted to create stereotypical views of groups, such as the genetic interpretation of intelligence test scores for racial groups (Hudson, 2015). Moreover, such stereotypes can lead to differential reactivity to the measurement procedure, as seen in stereotype threat and stereotype lift (Steele & Aronson, 1998; Walton & Cohen, 2003). More broadly, constructs that stem from a White dominant perspective could be biased against underrepresented minorities (American Psychological Association 2022; Horin et al., 2012; Shelton, 2000). As such, a holistic consideration of fairness and bias in conceptualization and measurement is warranted, whereby the consequences of constructs and their scores assumed to reflect them are interrogated.

CONCLUSIONS

All empirical research in social psychology requires the measurement of constructs. As such, the quality and impact of social-psychological research hinge in no small measure on effective measurement. Effective measurement begins and ends with well-delineated constructs. Such constructs specify attributes that are the focus of measurement strategies. Those strategies generate scores, which allow for description and hypothesis testing about the construct, its association with other constructs, and its relevance for policy and decision-making. This intimate connection between construct and measure in effective measurement is not always recognized, resulting in

measures only loosely connected to a construct and constructs for which no suitable measures have been developed. Measurement and construct theories summarized early in the chapter highlighted the connections between them and the need to always consider them in tandem.

The broad and inclusive coverage of targets, attributes, forms, and sources of information is a call for greater diversity in conceptualizing and measuring constructs in social psychology. The rationale for this call is threefold. First, no measure or measurement approach is without limiting assumptions and shortcomings. For that reason, description and hypothesis testing based on constructs always measured in the same way or using the same source of information is necessarily limited. Depending on the target, most constructs manifest multiple attributes (e.g., cognitive, affective), and most attributes can be measured using multiple sources of information. The robustness of inferences about constructs is strengthened by the use of multiple sources of information that capture different attributes of the construct. Second, the reliance on self-report measurement has contributed to a bias toward focusing on the person. Receiving relatively little attention are measures of groups, objects, situations, and environments. Measurement of constructs related to these targets in a manner that is not filtered through people's perceptions or limited by their exposure to them is challenging. Organic data relevant to these harder-to-measure targets are generated in large amounts and are often available to researchers. Increasingly inexpensive and accessible digital equipment is another means of obtaining objective information about targets other than people. Third, digital information generated as a byproduct of daily living in social settings suggests new constructs and novel manifestations of extant constructs while suggesting creative ways to measure them. As people increasingly engage in meaningful social interaction mediated by digital devices in digital contexts, a robust and relevant social psychology will include constructs and measures for describing and explaining it. In short, a broader view of measurement in social psychology points the way to more robust measurement in general and means of measuring constructs specified by theories beyond those filtered through the perception and experience of the person as reflected in self-reports.

As is true of all measurement approaches, self-report measurement is limited; however, for many constructs of interest to social psychologists, the self-report approach is a suitable means of generating scores. Moreover, for practical reasons (e.g., cost, reach), self-report measurement is likely to continue as a primary approach to measuring constructs in social psychology. For these reasons, the chapter devoted particular attention to the design and evaluation of self-report items, highlighting ways that self-report measurement could be conducted more thoughtfully with attention to rigor in the specification of all aspects of items. The design of items often focuses exclusively on the statements or questions to be rated, ignoring influential aspects of items, such as response options and labels. When self-report measurement is the preferred source of information about a construct, careful consideration of all aspects of item design that might influence the scores that will be treated as evidence of the construct is warranted.

In sum, measuring constructs is at the heart of research in social psychology. Although effective measurement does not ensure rigorous and reproducible research, it is essential for it; therefore, measurement must be done well. Doing so will require more attention to the delineation of constructs, more diversity and creativity in measuring them, and strong, clear connections between the products of measurement and the conceptualization of constructs. The result will be a more robust, informative, and influential social psychology.

AUTHOR NOTE

The chapter benefited from generous feedback and suggestions from Fernanda Andrade, Erin Davisson, Igor Grossman, and Jolynn Pek on a draft of the chapter.

REFERENCES

Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science: A call to action. *European Journal of Psychological Assessment, 38*(1), 1–5. https://doi.org/10.1027/1015-5759/a000699

Allport, G. (1935). Attitudes. In C. Murchison (Ed.), *Handbook of social psychology* (pp. 789–844). Clark University Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychological Association, APA Task Force on Guidelines for Assessment and Intervention with Persons with Disabilities. (2022). *Guidelines for assessment and intervention with persons with disabilities*. https://www.apa.org/about/policy/guidelines-assessment-intervention-disabilities.pdf

Andresen, E. M., Malmstrom, T. K., Schootman, M., Wolinsky, F. D., Miller, J. P., & Miller, D. K. (2013). Observer ratings of neighborhoods: Comparison of two methods. *BMC Public Health 13*, 1024. https://doi.org/10.1186/1471-2458-13-1024

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*(3), 666–678. https://doi.org/10.3758/s13428-011-0089-5

Baksh, R. A., Abrahams, S., Auyeung, B., & MacPherson, S. E. (2018). The Edinburgh Social Cognition Test (ESCoT): Examining the effects of age on a new measure of theory of mind and social norm understanding. *PLoS ONE 13*(4), e0195818. https://doi.org/10.1371/journal.pone.0195818

Balsa-Barreiro, J., Menendez, M., & Morales, A. J. (2022). Scale, context, and heterogeneity: The complexity of the social space. *Scientific Reports, 12*, 9037. https://doi.org/10.1038/s41598-022-12871-5

Bandura A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), Self-efficacy beliefs of adolescents (pp. 307–337). Information Age Publishing.

Bartels, M., Cacioppo, J. T., Hudziak, J. J., & Boomsma, D. I. (2008). Genetic and environmental contributions to stability in loneliness throughout childhood. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 147*(3), 385–391. https://doi.org/10.1002/ajmg.b.30608

Bartoszek, G., & Cervone, D. (2022). Measuring distinct emotional states implicitly: The role of response speed. *Emotion, 22*(5), 954–970. https://doi.org/10.1037/emo0000894

Basile, A. G., & Toplak, M. E. (2015). Four converging measures of temporal discounting and their relationships with intelligence, executive functions, thinking dispositions, and behavioral outcomes. *Frontiers in Psychology, 6*, 728. https://doi.org/10.3389/fpsyg.2015.00728

Baumeister, R. F., Muraven, M., & Tice, D. M. (2000). Ego depletion: A resource model of volition, self-regulation, and controlled processing. *Social Cognition, 18*(2), 130–150. https://doi.org/10.1521/soco.2000.18.2.130

Bayoumy, K., Gaber, M., Elshafeey, A., Mhaimeed, O., Dineen, E. H., Marvel, F. A., Martin, S. S., Muse, E. D., Turakhia, M. P., Tarakji, K. G., & Elshazly, M. B. (2021). Smart wearable devices in cardiovascular care: Where we are and how to move forward. *Nature Reviews Cardiology*, *18*(8), 581–599. https://doi.org/10.1038/s41569-021-00522-7

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing, *Public Opinion Quarterly, 71*(2), 287–311, https://doi.org/10.1093/poq/nfm006

Berntson, G. G., Bigger, T., Dwain, E.L., Grossman, P., Kaufmann, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, P., Stone, P. H., & van der Molen, M. W. (1997). Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology, 34*(6), 623–648. https://doi.org/10.1111/j.1469-8986.1997.tb02140.x

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Blanton, H., & Jaccard, J. (2018). From principles to measurement: Theory-based tips on writing better questions. In H. Blanton, J. M. LaCroix, & G. D. Webster (Eds.), *Measurement in social psychology* (pp. 1–25). Routledge.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42*(2), 192–212. https://doi.org/10.1016/j.jesp.2005.07.003

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology, 94*(3), 567–582. https://doi.org/10.1037/a0014665

Block J. (2000). Three tasks for personality psychology. In L. R. Bergman, R. B. Cairns, L. G., Nilsson, & L. Nystedt (Eds.), *Developmental science and the holistic approach* (pp. 155–164). Erlbaum.

Block, J., & Block, J. H. (1981). Studying situational dimensions: A grand perspective and some limited empiricism. In D. M. Magnusson (Ed.), *Toward a psychology of situations: An interactional perspective* (pp. 85–103). Erlbaum.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Boczkowski, P. J., & Mitchelstein, E. (2021). *The digital environment: How we live, learn, work, and play now*. MIT Press.

Boehnke, K. (2022). Let's compare apples and oranges! A plea to demystify measurement equivalence. *American Psychologist, 77*(9), 1160–1168. https://doi.org/10.1037/amp0001080

Boehnke, K., Arnaut, C., Bremer, T., Chinyemba, R., Kiewitt, Y., Koudadjey, A. K., Mwangase, R., & Neubert, L. (2014). Toward emically informed cross-cultural comparisons: A suggestion. *Journal of Cross-Cultural Psychology, 45*(10), 1655–1670. https://doi.org/10.1177/0022022114547571

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*, 579–616. https://doi.org/10.1146/annurev.psych.54.101601.145030

Bollen, K. A.., & Barb, K. H. (1981). Pearson's *R* and coarsely categorized measures. *American Sociological Review, 46*(2), 232–239. https://doi.org/10.2307/2094981

Bollen, K. A., & Hoyle, R. H. (1990). Perceived cohesion: A conceptual and empirical examination. *Social Forces*, *69*(2), 479–504. https://doi.org/10.1093/sf/69.2.479

Bollen, K. A., & Hoyle, R. H. (2023). Latent variables in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp. 97–109). Guilford Press.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314. https://doi.org/10.1037/0033-2909.110.2.305

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.) *Handbook of causal analysis for social research* (pp. 301–328). Springer. https://doi.org/10.1007/978-94-007-6094-3_15

Borgatta, E. F., & Bohrnstedt, G. W. (1974). Some limitations on generalizability from social psychological experiments. *Sociological Methods & Research, 3*(1), 111–120. https://doi.org/10.1177/004912417400300105

Borsboom, D. (2005). Measuring the mind: Conceptual issues in contemporary psychometrics. Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11, Suppl 3), S176–S181. https://doi.org/10.1097/01.mlr.0000245143.08679.cc

Borsboom, D. (2008). Latent variable theory. *Measurement*, *6*(1-2), 25–53. https://doi.org/10.1080/15366360802035497

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Borsboom, D., & Zand Scholten, A. (2008). The Rasch model and additive conjoint measurement theory from the perspective of psychometrics. *Theory & Psychology*, *18*(1), 111–117. https://doi.org/10.1177/0959354307086925

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., Dawson, M. E., & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology, 49*(8), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Bouwman, H., de Reuver, M., Heerschap, N., & Verkasalo, H. (2013). Opportunities and problems with automated data collection via smartphones. *Mobile Media & Communication, 1*(1), 63–68. https://doi.org/10.1177/2050157912464492

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, 27(3). 281–291, https://doi.org/10.1093/pubmed/fdi031

Bowling, N. A., Gibson, A. M. & DeSimone, J. A. (2022). Stop with the questions already! Does data quality suffer for scales positioned near the end of a lengthy questionnaire? *Journal of Business Psychology, 37*, 1099–1116. https://doi.org/10.1007/s10869-021-09787-8

Boyle, G. J., Saklofske, D. H., & Matthews, G. (Eds.) (2015). *Measures of personality and social psychological constructs*. Academic Press.

Brauer, M., Wasel, W., & Niedenthal, P. (2000). Implicit and explicit components of prejudice. *Review of General Psychology, 4*(1), 79–101. https://doi.org/10.1037/1089-2680.4.1.79

Brennan R. L. (2001). *Generalizability theory*. Springer.

Brenner, P. S., & DeLamater, J. (2016). Lies, damned lies, and survey self-reports? Identity as a cause of measurement bias. *Social Psychology Quarterly, 79*(4), 333–354. https://doi.org/10.1177/0190272516628298

Brewer, M., & Crano, W. (2014). Research design and issues of validity. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 11–26). Cambridge University Press. https://doi.org/10.1017/CBO9780511996481.005

Briggs, D. C. (2021). *Historical and conceptual foundations of measurement in the human sciences*. Routledge. https://doi.org/10.1201/9780429275326

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE*, *8*(4), e60188. https://doi.org/10.1371/journal.pone.0060188

Brislin, R. W. (1980) Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology: Methodology* (pp. 389–444). Allyn and Bacon.

Brown, N. A., Blake, A. B., & Sherman, R. A. (2017). A snapshot of the life as lived: Wearable cameras in social and personality psychological science. *Social Psychological and Personality Science, 8*(5), 592–600. https://doi.org/10.1177/1948550617703170

Brown-Iannuzzi, J. L., Lundberg, K. B., & McKee, S. (2017). The politics of socioeconomic status: How socioeconomic status may influence political attitudes and engagement. *Current Opinion in Psychology*, *18*, 11–14. https://doi.org/10.1016/j.copsyc.2017.06.018

Burt, C. H. (2022). Challenging the utility of polygenic scores for social science: Environmental confounding, downward causation, and unknown biology. *Behavioral and Brain Sciences, 46*, e207. https://doi.org/10.1017/s0140525x22001145

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Cacioppo, J. T., Berntson, G. G., & Decety, J. (2010). Social neuroscience and its relation to social psychology. *Social Cognition. 28*(6), 675–685. doi:10.1521/soco.2010.28.6.675

Cacioppo, J. T., Petty, R. E., & Tassinary, L. G. (1989). Social psychophysiology: A new look. In L. Berkowitz (Ed.), *Advances in experimental social psychology, 22*, 39–91. https://doi.org/10.1016/S0065-2601(08)60305-6

Cafarella, M. J., Madhavan, J., & Halevy, A. (2009). Web-scale extraction of structured data. *ACM SIGMOD Record*, *37*(4), 55–61. https://doi.org/10.1145/1519103.1519112

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Campbell, N. R. (1920). *Physics: The elements*. Cambridge University Press.

Carifio, J., & Perla, R. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences, 3*(3), 106–116. https://doi.org/10.3844/jssp.2007.106.116

Carpenter, N. C., Rangel, B., Jeon, G., & Cottrell, J. (2017). Are supervisors and coworkers likely to witness employee counterproductive work behavior? An investigation of observability and self-observer convergence. *Personnel Psychology, 70*(4), 843–889. https://doi.org/10.1111/peps.12210

Cathain, A., & Thomas, K. J. (2004). "Any other comments?" Open questions on questionnaires – a bane or a bonus to research? *BMC Medical Research Methodology, 4*, 25. https://doi.org/10.1186/1471-2288-4-25

Chachamovich, E., Fleck, M. P., & Power, M. (2009). Literacy affected ability to adequately discriminate among categories in multipoint Likert Scales. Journal of Clinical Epidemiology, 62(1), 37–46. https://doi.org/10.1016/j.jclinepi.2008.03.002

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. (2017). The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*, *20*(1), 3–31. https://doi.org/10.1177/1094428115617004

Champagne, F. A. (2010). Epigenetic influence of social experiences across the lifespan. *Developmental Psychobiology, 52*(4), 299–311. https://doi.org/10.1002/dev.20436

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83*(2), 234–246. https://doi.org/10.1037/0021-9010.83.2.234

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

Cho, E. (2021). Neither Cronbach's alpha nor McDonald's omega: A commentary on Sijtsma and Pfadt. *Psychometrika, 86*(4), 877–886. https://doi.org/10.1007/s11336-021-09801-1

Christensen, T. C., Feldman Barrett, L., Bliss-Moreau, E., Lebo, K., & Kaschub, C. (2003). A practical guide to experience-sampling procedures. *Journal of Happiness Studies*, *4*, 53–78. https://doi.org/10.1023/A:1023609306024

Cialdini, R. B. (2009). We have to break up. *Perspectives on Psychological Science, 4*(1), 5–6. https://doi.org/10.1111/j.1745-6924.2009.01091.x

Cliff, N. (1989). Ordinal consistency and ordinal true scores. *Psychometrika, 54*(1), 75–91. https://doi.org/10.1007/BF02294450

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*(3), 186–190. https://doi.org/10.1111/j.1467-9280.1992.tb00024.x

Clifton, A. B. W., Stahlmann, A. G., Hofmann, J., Chirico, A., Cadwallader, R., & Clifton, J. D. W. (2023). Improving scale equivalence by increasing access to scale-specific information. *Perspectives on Psychological Science, 18*(4), 843–853. https://doi.org/10.1177/17456916221119396

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods, 25*(3), 259–270. https://doi.org/10.1037/met0000236

Cloos, L., Ceulemans, E., & Kuppens, P. (2023). Development, validation, and comparison of self-report measures for positive and negative affect in intensive longitudinal research. *Psychological Assessment, 35*(3), 189–204. https://doi.org/10.1037/pas0001200

CloudResearch (n.d.). *A simple formula for predicting the time to complete a study on Mechanical Turk*. https://www.cloudresearch.com/resources/blog/a-simple-formula-for-predicting-the-time-to-complete-a-study-on-mechanical-turk/

Coie, J. D., & Dodge, K. A. (1988). Multiple sources of data on social behavior and social status in the school: A cross-age comparison. *Child Development, 59*(3), 815–829. https://doi.org/10.1111/j.1467-8624.1988.tb03237.x

Conner, T. S., & Mehl, M. R. (2015). Ambulatory assessment: Methods for studying everyday life. In R. Scott, S. Kosslyn, & N. Pinkerton (Eds.), *Emerging trends in the social and behavioral sciences* (pp. 1–15). Wiley

Connors, B. L., Rende, R., & Colton, T. J. (2016). Beyond self-report: Emerging methods for capturing individual differences in decision-making process. *Frontiers in Psychology, 7*, 312. https://doi.org/10.3389/fpsyg.2016.00312

Connors, G. J., & Maisto, S. A. (2003). Drinking reports from collateral individuals. *Addiction, 90*(s2), 21–29. https://doi.org/10.1046/j.1359-6357.2003.00585.x

Cording, M., Christmann, P., & Weigelt, C. (2010). Measuring theoretically complex constructs: The case of acquisition performance. *Strategic Organization, 8*(1), 11–41. https://doi.org/10.1177/1476127009355892

Cornet, V. P., & Holden, R. J. (2018). Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of Biomedical Informatics*, *77*, 120–132. https://doi.org/10.1016/j.jbi.2017.12.008

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences, 114*(38), E7900–E7909. https://doi.org/10.1073/pnas.1702247114

Craig, K., Hale, D., Grainger, C., & Stewart, M. E. (2020). Evaluating metacognitive self-reports: Systematic reviews of the value of self-report in metacognitive research. *Metacognition Learning 15*(2), 155–213. https://doi.org/10.1007/s11409-020-09222-y

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

Cronbach, L. J., & Gleser, C. G. (1957). *Psychological tests and personnel decisions*. University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418. https://doi.org/10.1177/0013164404266386

Csikszentmihalyi, M., Larson, R., & Prescott, S. (1977). The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, *6*(3), 281–294. https://doi.org/10.1007/BF02138940

Dalal, R. S., Alaybek, B., & Lievens, F. (2020). Within-person job performance variability over short timeframes: Theory, empirical research, and practice. *Annual Review of Organizational*

Psychology and Organizational Behavior, 7, 421–449. https://doi.org/10.1146/annurev-orgpsych-012119-045350

Dalege, J., Borsboom, D., van Harreveld, F., Lunansky, G., & van der Maas, H. L. J. (2018). The attitudinal entropy (AE) framework: Clarifications, extensions, and future directions. *Psychological Inquiry, 29*(4), 218–228. https://doi.org/10.1080/1047840X.2018.1542235

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The causal attitude network (CAN) model. *Psychological Review, 123*(1), 2–22. https://doi.org/10.1037/a0039802

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2017). Network analysis on attitudes: A brief tutorial. *Social Psychological and Personality Science*, *8*, 528–537. https://doi.org/10.1177/1948550617709827

Dalege, J., & van der Maas, H. L. J. (2020). Accurate by being noisy: A formal network model of implicit measures of attitudes. *Social Cognition*, *38*(Supplement), s26–s41. https://doi.org/10.1521/soco.2020.38.supp.s26

Das Swain, V., Saha, K., Rajvanshy, H., Sirigiri, A., Gregg, J. M., Lin, S., Martinez, G. J., Mattingly, S. M., Mirjafari, S., Mulukutla, R., Nepal, S., Nies, K., Reddy, M. D., Robles-Granda, P., Campbell, A.T., Chawla, N. V., D'Mello, S., Dey, A. K., Jiang, K. . . . De Choudhury, M. (2019). A multisensor person-centered approach to understand the role of daily activities in job performance with organizational personas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(4), 1–27. https://doi.org/10.1145/3369828

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

Davidson, B., Ellis, D., Stachl, C., Taylor, P., & Joinson, A. (2022). Measurement practices exacerbate the generalizability crisis: Novel digital measures can help. *Behavioral and Brain Sciences, 45*, e10. https://doi.org/10.1017/S0140525X21000534

De Beuckelaer, A., & Swinnen, G. (2011). Biased latent variable mean comparisons due to measurement noninvariance: A simulation study. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 117–147). Routledge/Taylor & Francis Group.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368. https://doi.org/10.1037/a0014211

De Boeck, P., Pek, J., Walton, K., Wegener, D. T., Turner, B., Andersen, B., Beauchaine, T. P., Lecavalier, L., Myung, J. I., & Petty, R. E. (2023). Questioning psychological constructs: Current issues and proposed changes. *Psychological Inquiry, 34*(4), 239–257. https://doi.org/10.1080/1047840X.2023.2274429

Desjardins, J. (2019, March 13). *What happens in an Internet minute in 2019?* Visual Capitalist. https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/

Diener, E., Thapa, S., & Tay, L. (2020). Positive emotions at work. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*(1), 451–477. https://doi.org/10.1146/annurev-orgpsych-012119-044908

Downs, A. C. (1990). The social biological constructs of social competency. In T. P. Gullotta, G. R. Adams, & R. Montemayor (Eds.), *Developing social competency in adolescence* (pp. 43–94). Sage Publications.

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (Grit-S). *Journal of Personality Assessment, 91*(2), 166–174

Dufner, M., & Krause, S. (2023). On how to be liked in first encounters: The effects of agentic and communal behaviors on popularity and unique liking. *Psychological Science, 34*(4), 481–489. https://doi.org/10.1177/09567976221147258

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

Dwyer, R. J., Kushlev, K, & Dunn, E. W. (2018). Smartphone use undermines enjoyment of face-to-face social interactions. *Journal of Experimental Social Psychology, 78*, 233–239. https://doi.org/10.1016/j.jesp.2017.10.007

Ebner-Priemer, U. W., Eid, M., Kleindienst, N., Stabenow, S., & Trull, T. J. (2009). Analytic strategies for understanding affective (in)stability and other dynamic processes in psychopathology. *Journal of Abnormal Psychology*, *118*(1), 195–202. https://doi.org/10.1037/a0014868

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*, 155–174. https://doi.org/10.1037/1082-989x.5.2.155

Eid, M., & Diener, E. (2006). Introduction: The need for multimethod measurement in psychology. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 3–8). American Psychological Association. https://doi.org/10.1037/11383-001

Eid, M., & Diener, E. (2009). Norms for experiencing emotions in different cultures: Inter- and intranational differences. In E. Diener (Ed.), *Culture and well-being* (pp. 169–202). Springer. https://doi.org/10.1007/978-90-481-2352-0_9

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151. https://doi.org/10.1177/1073191120957102

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences, 113*(28), 7900–7905. https://doi.org/10.1073/pnas.1602413113

Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika, 62*(4), 495–523. https://doi.org/10.1007/BF02294640

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Ewbank, D. C. (2008). Biomarkers in social science research on health and aging: A review of theory and practice. In M. Weinstein, J. W. Vaupel, & K. W. Wachter (Eds.), *Biosocial surveys* (pp. 156–171). National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK62447/

Farias, S. T., Mungas, D., Reed, B. R., Cahn-Weiner, D., Jagust, W., Baynes, K., & Decarli, C. (2008). The measurement of everyday cognition (ECog): Scale development and psychometric properties. *Neuropsychology, 22*(4), 531–544. https://doi.org/10.1037/0894-4105.22.4.531

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M. S. Clark (Eds.), *Research methods in personality and social psychology* (pp. 74–97). Sage Publications.

Fischer, H. (2011). *A history of the central limit theorem: From classical to modern probability theory*. Springer.

Fisher, C. D., & To, M. L. (2012). Using experience sampling methodology in organizational behavior. *Journal of Organizational Behavior*, *33*(7), 865–877. https://doi.org/10.1002/job.1803

Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of Occupational Health Psychology, 21*(1), 3–23. https://doi.org/10.1037/a0039139

Fishman, J., Yang, C. & Mandell, D. (2021). Attitude theory and measurement in implementation science: A secondary review of empirical studies and opportunities for advancement. *Implementation Science, 16*(1), 87. https://doi.org/10.1186/s13012-021-01153-9

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fleeson, W., & Gallagher, P. (2009). The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of Personality and Social Psychology*, *97*(6), 1097–1114. https://doi.org/10.1037/a0016786

Fleeson, W., & Law, M. K. (2015). Trait enactments as density distributions: The role of actors, situations, and observers in explaining stability and variability. Journal of Personality and Social Psychology, 109(6), 1090–1104. https://doi.org/10.1037/a0039517.

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science, 3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Foltz, C., Morse, J. Q., Calvo, N., & Barber, J. P. (1997). Self- and observer ratings on the NEO-FFI in couples: Initial evidence of the psychometric properties of an observer form. *Assessment, 4*(3), 287–295. https://doi.org/10.1177/107319119700400308

Friborg, O., & Rosenvinge, J. H. (2013). A comparison of open-ended and closed questions in the prediction of mental health. *Quality & Quantity, 47*(3), 1397–1411. https://doi.org/10.1007/s11135-011-9597-8

Fried, E. I. (2017). What are psychological constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health Psychology Review, 11*(2), 130–134. https://doi.org/10.1080/17437199.2017.1306718

Frost, B. C., Ko, C.-H. E., & James, L. R. (2007). Implicit and explicit personality: A test of a channeling hypothesis for aggressive behavior. *Journal of Applied Psychology, 92*(5), 1299–1319. https://doi.org/10.1037/0021-9010.92.5.1299

Funder, D. C., Guillaume, E., Kumagai, S., Kawamoto, S., & Sato, T. (2012). The person-situation debate and the assessment of situations. *Japanese Journal of Personality, 21*(1), 1–11. https://doi.org/10.2132/personality.21.1

Geiser, C., Götz, T., Preckel, F., & Freund, P. A. (2017). States and traits: Theories, models, and assessment. *European Journal of Psychological Assessment, 33*(4), 219-223. https://doi.org/10.1027/1015-5759/a000413

Gergen, K. J. (2011). The self as social construction. *Psychological Studies, 56*(1), 108–116. https://doi.org/10.1007/s12646-011-0066-1

Gnambs, T., & Buntins, K. (2017). The measurement of variability and change in life satisfaction: A comparison of single-item and multi-item instruments. *European Journal of Psychological Assessment, 33*(4), 224–238. https://doi.org/10.1027/1015-5759/a000414

Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods, 47*, 1237–1259. https://doi.org/10.3758/s13428-014-0533-4

Gosling, S. D., Augustine, A. A., Vazire, S., Holtzman, N., & Gaddis, S. (2011). Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information. *CyberPsychology, Behavior, and Social Networking, 14*(9), 483–488. https://doi.org/10.1089/cyber.2010.0087

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of Personality and Social Psychology, 74*(5), 1337–1349. https://doi.org/10.1037//0022-3514.74.5.1337

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*(4), 827–838. https://doi.org/10.1177/001316447703700403

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist*, *52*(10), 1115–1124. https://doi.org/10.1037/0003-066X.52.10.1115

Greenfield, T. K., Bond, J., & Kerr, W. C. (2014). Biomonitoring for improving alcohol consumption surveys: The new gold standard? *Alcohol Research: Current Reviews*, *36*(1), 39–45.

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, *7*(2), 99–108. https://doi.org/10.1177/1745691611434210

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5), 861–871. https://doi.org/10.1093/poq/nfr057

Grysman, A. (2015). Collecting narrative data on Amazon's Mechanical Turk. *Applied Cognitive Psychology, 29*(4), 573–583. https://doi.org/10.1002/acp.3140

Guttman L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255–282. https://doi.org/10.1007/BF02288892.

Guttman, L. (1971). Measurement as structural theory. *Psychometrika*, *3*(4), 329–347. https://doi.org/10.1007/BF02291362

Haddock, G., & Zanna, M. P. (1998). On the use of open-ended measures to assess attitudinal components. *British Journal of Social Psychology, 37*(2), 129–149. https://doi.org/10.1111/j.2044-8309.1998.tb01161.x

Hadwin, A. F., Winne, P. H., Stockley, D. B., Nesbit, J. C., & Woszczyna, C. (2001). Context moderates students' self-reports about how they study. *Journal of Educational Psychology, 93*(3), 477–487. https://doi.org/10.1037/0022-0663.93.3.477

Haeffel, G. J., & Howard, G. S. (2010). Self-report: Psychology's four-letter word. *American Journal of Psychology, 123*(2), 181–188. https://doi.org/10.5406/amerjpsyc.123.2.0181

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*(3), 1369–1392. https://doi.org/10.1037/Fa0035028

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. M. (2007). The integrated state-trait model. *Journal of Research in Personality, 41*(2), 295–315. https://doi.org/10.1016/j.jrp.2006.04.003

Han, K., Colarelli, S. M., & Weed, N. C. (2019). Methodological and statistical advances in the consideration of cultural diversity in assessment: A critical review of group classification and measurement invariance testing. *Psychological Assessment, 31*(12), 1481–1496. https://doi.org/10.1037/pas0000731

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016) Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives in Psychological Science, 11*(6), 838–854. https://doi.org/10.1177/1745691616650285

Harkness, J. A. (2007). Improving the comparability of translations. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (pp. 79–93). Sage Publications. https://doi.org/10.4135/9781849209458.n4

Harmon-Jones, C., Bastian, B., & Harmon-Jones, E. (2016) The Discrete Emotions Questionnaire: A new tool for measuring state self-reported emotions. *PLoS ONE 11*(8), e0159915. https://doi.org/10.1371/journal.pone.0159915

Harrewijn, A., van der Molen, M. J. W., van Vliet, I. M., Tissier, R. L. M., & Westenberg, P. M. (2018). Behavioral and EEG responses to social evaluation: A two-generation family study on social anxiety. *NeuroImage: Clinical, 17*, 549–562. https://doi.org/10.1016/j.nicl.2017.11.010

Harrison, L. D. (1995). The validity of self-reported data on drug use. *Journal of Drug Issues, 25*(1), 91–111. https://doi.org/10.1177/002204269502500107

Haslbeck, J. M. B., Epskamp, S., Marsman, M., & Waldorp, L. J. (2021). Interpreting the Ising Model: The input matters. *Multivariate Behavioral Research*, *56*(2), 303–313. https://doi.org/10.1080/00273171.2020.1730150

Haws, K. L., Davis, S. W., & Dholakia, U. M. (2016). Control over what? Individual differences in general versus eating and spending self-control. *Journal of Public Policy & Marketing, 35*(1), 37–57. https://doi.org/10.1509/jppm.14.149

Heainisch, D. A., & Jex, S. M. (1998). Measurement of negative affectivity: A comparison of self-reports and observer ratings. *Work & Stress, 12*(2), 145–160. https://doi.org/10.1080/02678379808256856

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83. https://doi.org/10.1017/S0140525X0999152X

Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry, 19*(2), 49–71. https://doi.org/10.1080/10478400802049936

Herreen, D., & Zajac, I. T. (2017). The reliability and validity of a self-report measure of cognitive abilities in older adults: More personality than cognitive function. *Journal of Intelligence, 6*(1),

1. https://doi.org/10.3390/jintelligence6010001

Hirsch, P., Nolden, S., Declerck, M., & Koch, I. (2018). Common cognitive control processes underlying performance in task-switching and dual-task contexts. *Advances in Cognitive Psychology, 14*(3), 62–74. https://doi.org/10.5709/acp-0239-y

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science, 355*(6324), 486–488. https://doi.org/10.1126/science.aal3856

Höhne, J. K., Krebs, D., & Kühnel, S.-M. (2021). Measurement properties of completely and end labeled unipolar and bipolar scales in Likert-type questions on income (in)equality. *Social Science Research, 97*, 102544. https://doi.org/10.1016/j.ssresearch.2021.102544

Höhne, J. K., Krebs, D., & Kühnel, S.-M. (2022). Measuring income (in)equality: Comparing survey questions with unipolar and bipolar scales in a probability-based online panel. *Social Science Computer Review, 40*(1), 108–123. https://doi.org/10.1177/0894439320902461

Horin, E. V., Hernandez, B., & Donoso, O. A. (2012). Behind closed doors: Assessing individuals from diverse backgrounds. *Journal of Vocational Rehabilitation, 37*(2), 87–97. https://doi.org/10.3233/JVR-2012-0602

Horvath, F. W. (1982). Forgotten unemployment: Recall bias in retrospective data. *Monthly Labour Review, 150*(3), 40–44. https://stats.bls.gov/opub/mlr/1982/03/rpt3full.pdf

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology, 64*(2), 144–150. https://doi.org/10.1037/0021-9010.64.2.144

Hoyle, R. H., Lynam, D. R., Miller, J. D., & Pek, J. (2023). The questionable practice of partialing to refine scores on and inferences about measures of psychological constructs. *Annual Review of Clinical Psychology, 19*, 155–176. https://doi.org/10.1146/annurev-clinpsy- 071720-015436

Hoyle, R. H., Stephenson, M. T., Palmgreen, P., Lorch, E. P., & Donohew, R. L. (2002). Reliability and validity of a brief measure of sensation seeking. *Personality and Individual Differences*, *32*(3), 401–414. https://doi.org/10.1016/S0191-8869(01)00032-0

Hudson, J. B. (2015). Scientific racism: The politics of tests, race and genetics. *The Black Scholar*, *25*(1), 3–10. https://doi.org/10.1080/00064246.1995.11430694

Hyatt, C. S., Sleep, C. E., Lamkin, J., Maples-Keller, J. L., Sedikides, C., Campbell, W. K., & Miller, J. D. (2018). Narcissism and self-esteem: A nomological network analysis. *PLoS One, 13*(8), e0201088. https://doi.org/10.1371/journal.pone.0201088

Jebb, A. T., Ng, V., & Tay, L. (2021). A review of key Likert scale development advances: 1995–2019. *Frontiers in Psychology, 12*, 637547. https://doi.org/10.3389/fpsyg.2021.637547

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61*(4), 521–551. https://doi.org/10.1111/j.1467-6494.1993.tb00781.x

Jones, J. T., Pelham, B. W., Mirenberg, M. C., & Hetts, J. J. (2002). Name letter preferences are not merely mere exposure: Implicit egotism as self-regulation. *Journal of Experimental Social Psychology, 38*(2), 170–177. https://doi.org/10.1006/jesp.2001.1497

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133. https://doi.org/10.1007/BF02291393

Kane, M. T., (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. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kaplan, K. J. (1972). On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological Bulletin, 77*(5), 361–372. https://doi.org/10.1037/h0032590

Kelley, H. H., Holmes, J. G., Kerr, N. L., Reis, H. T., Rusbult, C. E., & van Lange, P. A. M. (2003). *An atlas of interpersonal situations*. Cambridge University Press.

Kelley, T. L. (1927). *Interpretation of educational measurements*. Macmillan.

Kenny, D. A., & La Voie, L. (1984). The social relations model. *Advances in Experimental Social Psychology, 18*, 141–182. https://doi.org/10.1016/S0065-2601(08)60144-6

Kernis, M. H., Cornell, D. P., Sun, C. R., Berry, A., Harlow, T. (1993). There's more to self-esteem than whether it is high or low: The importance of stability of self-esteem. *Journal of Personality and Social Psychology, 65*(6), 1190–1204. https://doi.org/10.1037//0022-3514.65.6.1190

Keusch, F., Struminskaya, B., Kreuter, F., & Weichbold, M. (2021). Combining active and passive mobile data collection: A survey of concerns. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Big data meets survey science: A collection of innovative methods* (pp. 657–682). Wiley. https://doi.org/10.1002/9781118976357.ch22

Kitayama, S., & Uskul, A. K. (2011). Culture, mind, and the brain: Current evidence and future directions. *Annual Review of Psychology, 62*, 419–449. https://doi.org/10.1146/annurev-psych-120709-145357

Klein, H., Springfield, C. R., & Pinkham, A. E. (2024). Measuring social cognition within the university: The Social Cognition Psychometric Evaluation (SCOPE) battery in an undergraduate sample. *Applied Neuropsychology: Adult*, 31(5), 866–873. https://doi.org/10.1080/23279095.2022.2082875

Kluger, A. N., Malloy, T. E., Pery, S., Itzchakov, G., Castro, D. R., Lipetz, L., Sela, Y., Turjeman-Levi, Y., Lehmann, M., New, M., & Borut, L. (2021). Dyadic listening in teams: Social relations model. *Applied Psychology, 70*(3), 1045–1099. https://doi.org/10.1111/apps.12263

Koczkodaj, W.W., Kakiashvili, T., Szymańska, A., Montero-Marin, J., Araya, R., Garcia-Campayo, J., Rutkowski, K., & Strzałka, D. (2017). How to reduce the number of rating scale items without predictability loss? *Scientometrics, 111*, 581-593. https://doi.org/10.1007/s11192-017-2283-4

Kornbrot, D. E., Wiseman, R., & Georgiou, G. J. (2018). Quality science from quality measurement: The role of measurement type with respect to replication and effect size magnitude in psychological research. *PLoS ONE, 13*(2), e0192808. https://doi.org/10.1371/journal.pone.0192808

Kosinski, M. (2024). Using big data. In D. T. Gilbert, S. T. Fiske, E. J. Finkel, & W. B. Mendes (Eds.), *The Handbook of Social Psychology* (6th ed.). Situational Press.

Kotov, R., Krueger, R. F., Watson, D., Cicero, D. C., Conway, C. C., DeYoung, C. G., Eaton, N. R., Forbes, M. K., Hallquist, M. N., Latzman, R. D., Mullins-Sweatt, S. N., Ruggero, C. J., Simms, L. J., Waldman, I. D., Waszczuk, M. A., & Wright, A. G. C. (2021). The hierarchical taxonomy of psychopathology (HiTOP): A quantitative nosology based on consensus of evidence. *Annual Review of Clinical Psychology, 17*, 83–108. https://doi.org/10.1146/annurev-clinpsy-081219-093304

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Volume I: Additive and polynomial representations*. Academic Press.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1989). *Foundations of measurement, Volume II: Geometrical, threshold, and probabilistic representations*. Academic Press.

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., & Trappmann, M. (2020). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review, 38*(5), 533–549. https://doi.org/10.1177/0894439318816389

Kripke, S. A. (1980). *Naming and necessity*. Harvard University Press.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141–164). Wiley.

Kubiszyn, T., & Borich, G. D. (2016). *Educational testing and measurement*. Wiley.

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*(6), 1042–1060. https://doi.org/10.1037/a0020962

Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, *18*(1), 89–109. https://doi.org/10.1177/0959354307086924

Larsen, R. J., & Fredrickson, B. L. (1999). Measurement issues in emotion research. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: Foundations of hedonic psychology* (pp. 40–60). Russell Sage.

Larson, J. J., Whitton, S. W., Hauser, S.T., & Allen, J. P. (2007). Being close and being social: Peer ratings of distinct aspects of young adult social competence. *Journal of Personality Assessment, 89*(2), 136–148. https://doi.org/10.1080%2F00223890701468501

Latané, B. (1981). The psychology of social impact. *American Psychologist, 36*(4), 343–356. https://doi.org/10.1037/0003-066X.36.4.343

Lawson, K. M., & Robins, R. W. (2021). Sibling constructs: What are they, why do they matter, and how should you handle them? *Personality and Social Psychology Review, 25*(4), 344–366. https://doi.org/10.1177/10888683211047101

Leary, M. R. (2012). Sociometer theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 151–159). Sage Publications. https://doi.org/10.4135/9781446249222.n33

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., De Roover, K., Jak, S., Meitinger, K., Menold, N., Muthén, B., Rudnev, Schmidt, M. P., & van de Schoot, R. (2023). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research, 110*, 102805. https://doi.org/10.1016/j.ssresearch.2022.102805

Levy. J,, Goldstein, A., Influs, M., Masalha, S., Zagoory-Sharon, O., & Feldman, R. (2016). Adolescents growing up amidst intractable conflict attenuate brain response to pain of outgroup. *Proceedings of the National Academy of Sciences, 113*(48), 13696–13701. https://doi.org/10.1073/pnas.1612903113

Lewis, (2019). Comparison of four TAM item formats: Effect of response option labels and order. *Journal of Usability Studies, 14*(4), 224–236. https://uxpajournal.org/tam-formats-effect-response-labels-order/

Lieberz, J., Shamay-Tsoory, S. G., Saporta, N., Kanterman, A., Gorni, J., Esser, T., Kuskova, E., Schultz, J., Hurlemann, R., & Scheele, D. (2022). Behavioral and neural dissociation of social anxiety and loneliness. *Journal of Neuroscience, 42*(12), 2570–2583. https://doi.org/10.1523/JNEUROSCI.2029-21.2022

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 5–55. https://archive.org/details/likert-1932/

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology, 61*(4), 281–288. https://doi.org/10.1037/cap0000236

Linney, J. A. (2000). Assessing ecological constructs and community context. In J. Rappaport & E. Seidman (Eds.), *Handbook of community psychology* (pp. 647–668). Springer. https://doi.org/10.1007/978-1-4615-4193-6_27

Lira, B., O'Brien, J. M., Peña, P. A., Galla, B. M., D'Mello, S. Yeager, D. S., Defnet., A., Kautz, T., Munkacsy, K., & Duckworth, A. L. (2022). Large studies reveal how reference bias limits policy applications of self-report measures. *Scientific Reports, 12*, 19189. https://doi.org/10.1038/s41598-022-23373-9

Liu, H., Xie, Q. W., Lou, V. W. Q. (2019). Everyday social interactions and intra-individual variability in affect: A systematic review and meta-analysis of ecological momentary assessment studies. *Motivation and Emotion, 43*, 339–353. https://doi.org/10.1007/s11031-018-9735-x

Liu, Y., Nour, M. M., Schuck, N. W., Behrens, T. E. J., & Dolan, R. J. (2022). Decoding cognition from spontaneous neural activity. *Nature Reviews Neuroscience, 23*(4), 204–214. https://doi.org/10.1038/s41583-022-00570-z

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement, Volume III: Representation, axiomatization, and invariance*. Academic Press

Maitland, A. (2009). Should I label all scale points or just the end points for attitudinal questions? *Survey Practice, 2*(4). https://doi.org/10.29115/SP-2009-0014.

Marengo, D., Montag, C., Mignogna, A., & Settanni, M. (2022). Mining digital traces of Facebook activity for the prediction of individual differences in tendencies toward social networks use disorder: A machine learning approach. *Frontiers in Psychology, 13*, 830120. https:/doi.org/10.3389/fpsyg.2022.830120

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences: Developing a shared concept system for measurement*. Springer.

Markus, K. A., & Borsboom, D. (2012). The cat came back: Evaluating arguments against psychological measurement. *Theory & Psychology, 22*(4), 452–466. https://doi.org/10.1177/0959354310381155

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

Marsh, H. W. (1986). Global self-esteem: Its relation to specific facets of self-concept and their importance. *Journal of Personality and Social Psychology, 51*(6), 1224–1236. https://doi.org/10.1037/0022-3514.51.6.1224

Marsh, H. W. (1990). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review, 2*(2), 77–172. https://doi.org/10.1007/BF01322177

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., Maas, H. L. J. v. d., & Maris, G. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate Behavioral Research, 53*(1), 15–35. https://doi.org/10.1080/00273171.2017.1379379

Martin, G., Gavine, A., Inchley, J., & Currie, C. (2017). Conceptualizing, measuring and evaluating constructs of the adolescent neighbourhood social environment: A systematic review. *SSM-Population Health, 3*, 335–351. https://doi.org/10.1016/j.ssmph.2017.03.002

Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition, 1*(1), 16, 1–23. https://doi.org/10.5334/joc.18

Mauss, I. B., Levenson, R. W. McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion, 5*(2), 175–190. https://doi.org/10.1037/1528-3542.5.2.175

Mauss, I. B., & Robinson, M. D. (2009) Measures of emotion: A review. *Cognition and Emotion, 23*(2), 209–237. https://doi.org/10.1080/02699930802204677

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology, 67*(3), 267–277. https://doi.org/10.1016/j.jclinepi.2013.08.015

McDonald, R. P. (1999). *Test theory: A unified treatment*. Erlbaum.

McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, *49*(3), 212–230.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*(2), 300–307. https://doi.org/10.1037/0033-2909.115.2.300

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*(3), 293–299. https://doi.org/10.1037/1082-989X.1.3.293

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89*(3), 575–588. https://doi.org/10.1037/0033-2909.89.3.575

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). American Council on Education and National Council on Measurement in Education.

Meta Business Help Center (n.d.). *How Facebook distributes content*. https://www.facebook.com/business/help/718033381901819

Meyerson, W., Fineberg, S., Song, Y. K., Faber, A., Ash, G., Andrade, F. C., Corlett, P., Gerstein, M., & Hoyle, R. H. (2023). Estimation of bedtimes of Reddit users: Integrated analysis of timestamps and surveys. *JMIR Formative Research, 7*, e38112. https://doi.org/10.2196/38112

Michell, J. (1993). The origins of the representational theory of measurement: Helmholtz, Hölder, and Russell. *Studies in History and Philosophy of Science Part A*, *24*(2), 185–206. https://doi.org/10.1016/0039-3681(93)90045-L

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press.

Monahan, P. O., Alder, C. A., Khan, B. A., Stump, T., Boustani, M. A. (2014) The Healthy Aging Brain Care (HABC) Monitor: Validation of the patient self-report version of the clinical tool designed to measure and monitor cognitive, functional, and psychological health. *Clinical Interventions in Aging, 9*, 2123–2132. https://doi.org/10.2147/cia.s64140

Montag, C., Dagum, P., Hall, B. J., & Elhai, J. D. (2022). Do we still need psychological self-report questionnaires in the age of the Internet of Things? *Discover Psychology, 2*(1). https://doi.org/10.1007/s44202-021-00012-4

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*(4), 379–387. https://doi.org/10.1037/0003-066X.38.4.379

Moreau, D., & Wiebels, K. (2022). Psychological constructs as local optima. *Nature Reviews Psychology, 1*, 188–189-. https://doi.org/10.1038/s44159-022-00042-2

Moshe, I., Terhorst, Y., Opoku Asare, K., Sander, L. B., Ferreira, D., Baumeister, H., Mohr, D. C., & Pulkki-Råback, L. (2021). Predicting symptoms of depression and anxiety using smartphone and wearable data. *Frontiers in Psychiatry, 12*, 625247. https://doi.org/10.3389/fpsyt.2021.625247

Mulligan, K., Grant, J., Mockabee, S., & Monson, J. (2003). Response latency methodology for survey research: Measurement and modeling strategies. *Political Analysis, 11*(3), 289–301. https://doi.org/10.1093/pan/mpg004

Muscatell, K. A. (2021). Social psychoneuroimmunology: Understanding bidirectional links between social experiences and the immune system. *Brain, Behavior, and Immunity, 93*, 1–3. https://doi.org/10.1016/j.bbi.2020.12.023

Nagel, E., & Hempel, C. G. (1931). Measurement. *Erkenntnis, 2*, 313–335. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071456679&partnerID=40&md5=e46ad1ca7e8a2b850cf06def47219ea9

Nassar-McMillan, S. C., Wyer, M., Oliver-Hoyo, M., & Ryder-Burge, A. (2010). Using focus groups in preliminary instrument development: Expected and unexpected lessons learned. *The Qualitative Report, 15*(6), 1629–1642. https://doi.org/10.46743/2160-3715/2010.1368

Nelson, B. W., & Allen, N. B. (2018). Extending the passive-sensing toolbox: Using smart-home technology in psychological science. *Perspectives on Psychological Science, 13*(6), 718–733. https://doi.org/10.1177/1745691618776008

Newton, P. E., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Sage Publications.

Nichols, M., & Newsome, W. (1999). The neurobiology of cognition. *Nature, 402* (Suppl 6761), C35–C38. https://doi.org/10.1038/35011531

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259. http://dx.doi.org/10.1037/0033-295X.84.3.231

Nosek, B. A., & Greenwald, A. G. (2009). (Part of) the case for a pragmatic approach to validity: Comment on De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009). *Psychological Bulletin, 135*(3), 373–376. https://doi.org/10.1037/a0015047

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Science, 15*(4), 152–159. https://doi.org/10.1016/j.tics.2011.01.005

Nowak, A., Szamrej, J., & Latané, B. (1990). From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review. 97*(3), 362–376. doi:10.1037/0033-295X.97.3.362.

Odgers, C. L., Caspi, A., Bates, C. J., Sampson, R. J. & Moffitt, T. E. (2012). Systematic social observation of children's neighborhoods using Google Street View: A reliable and cost-effective method. *Journal of Child Psychology and Psychiatry, 53*(10), 1009–1017. https://doi.org/10.1111/j.1469-7610.2012.02565.x

Olderbak, S., & Wilhelm, O. (2020). Overarching principles for the organization of socioemotional constructs. *Current Directions in Psychological Science, 29*(1), 63–70. https://doi.org/10.1177/0963721419884317

Onoshima, T., Shiina, K., Ueda, T., & Kubo, S. (2019). Decline of Pearson's *r* with categorization of variables: A large-scale simulation. *Behaviormetrika, 46*, 389–399. https://doi.org/10.1007/s41237-019-00089-1

Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel I. (2020). Detection of bots in social media: A systematic review. *Information Processing & Management, 57*(4), 102250. https://doi.org/10.1016/j.ipm.2020.102250]

Orehek, E., & Human, L. J. (2017). Self-expression on social media: Do tweets present accurate and positive portraits of impulsivity, self-esteem, and attachment style? *Personality and Social Psychology Bulletin, 43*(1), 60–70. https://doi.org/10.1177/0146167216675332

Øverup, C. S., Brunson, J. A., & Mehta, P. D. (2021). A social relations model of need supportiveness. *Journal of Research in Personality, 94*, 104142. https://doi.org/10.1016/j.jrp.2021.104142

Pang, L., Li, H., Liu, Q., Luo, Y.-J., Mobbs, D., & Wu, H. (2022). Resting-state functional connectivity of social brain regions predicts motivated dishonesty. *NeuroImage, 256*, 119253. https://doi.org/10.1016/j.neuroimage.2022.119253

Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behavior, 5*, 1535–1547. https://doi.org/10.1038/s41562-021-01117-5

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). Guilford Press.

Paunonen, S. V., & Kam, C. (2014). The accuracy of roommate ratings of behaviors versus beliefs. *Journal of Research in Personality, 52*, 55–67. https://doi.org/10.1016/j.jrp.2014.07.006

Payne, B. K., Brown-Iannuzzi, J. L., & Hannay, J. W. (2017). Economic inequality increases risk taking. *Proceedings of the National Academy of Sciences, 114*(18), 4643–4648. https://doi.org/10.1073/pnas.1616453114

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*(4), 233–248. https://doi.org/10.1080/1047840X.2017.1335568

Pelham, B. W., Koole, S. L., Hardin, C. D., Hetts, J. J., Seah, E., & DeHart, T. (2005). Gender moderates the relation between implicit and explicit self-esteem. *Journal of Experimental Social Psychology, 41*(1), 84–89. https://doi.org/10.1016/j.jesp.2003.10.008

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237–255. https://doi.org/10.1177/014662167900300213

Perugini, M. (2005). Predictive models of implicit and explicit attitudes. *British Journal of Social Psychology, 44*(1), 29–45. https://doi.org/10.1348/014466604X23491

Peters, G.-J. Y., & Crutzen, R. (2017). Pragmatic nihilism: How a theory of nothing can help health psychology progress. *Health Psychology Review, 11*(2), 103–121. https://doi.org/10.1080/17437199.2017.1284015

Petitmengin, C., Remillieux, A., Cahour, B., & Carter-Thomas, S. (2013). A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes. *Consciousness and Cognition, 22*(2), 654–669. https://doi.org/10.1016/j.concog.2013.02.004

Phua, D. Y., & Christopoulos, G. (2014). Social neuroscience tasks: Employing fMRI to understand the social mind. In T. D. Papageorgiou, G. I. Christopoulos, & S. M. Smirnakis (Eds.), *Advanced brain neuroimaging topics in health and disease: Methods and applications* (pp. 653–678). InTechOpen Limited. https://doi.org/10.5772/58279

Podschuweit, N. (2021). How ethical challenges of covert observations can be met in practice. *Research Ethics, 17*(3), 309–327. https://doi.org/10.1177/17470161211008218

Polit, D. F. (2014). Getting serious about test-retest reliability: A critique of retest research and some recommendations. *Quality of Life Research, 23*(6), 1713–1720. https://doi.org/10.1007/s11136-014-0632-9

Popper, K. (1959). *The logic of scientific discovery*. Hutchinson.

Principles for the Validation and Use of Personnel Selection Procedures (2018). *Industrial and Organizational Psychology: Perspectives on Science and Practice, 11*(Supl 1), 2–97. https://doi.org/10.1017/iop.2018.195

Quirin, M., & Bode, R. C. (2014). An alternative to self-reports of trait and state affect: The Implicit Positive and Negative Affect Test (IPANAT). *European Journal of Psychological Assessment, 30*(3), 231–237. https://doi.org/10.1027/1015-5759/a000190.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Institut.

Rast, P., Zimprich, D., Van Boxtel, M., & Jolles, J. (2009). Factor structure and measurement invariance of the Cognitive Failures Questionnaire across the adult life span. *Assessment, 16*(2), 145–158. https://doi.org/10.1177/1073191108324440

Rauthmann, J. F., & Sherman, R. A. (2021). Conceptualizing and measuring the psychological situation. In D. Wood, S. J. Read, P. D. Harms, & A. Slaughter (Eds), *Measuring and modeling persons and situations* (pp. 427-463). Academic Press. https://doi.org/10.1016/B978-0-12-819200-9.00009-0

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173–184. https://doi.org/10.1177/01466216970212006

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement, 79*(1), 200–210. https://doi.org/10.1177/0013164417725127

Reis, H. T. (2008). Reinvigorating the concept of situation in social psychology. *Personality and Social Psychology Review, 12*(4), 311–329. https://doi.org/10.1177/1088868308321721

Rentfrow, P. J. (2010). Statewide differences in personality: Toward a psychological geography of the United States. *American Psychologist, 65*(6), 548–558. https://doi.org/10.1037/a0018194

Revelle, W., & Zinbarg, R. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods, 25*(1), 30–45. https://doi.org/10.1037/met0000220

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 27*(2), 151–161. https://doi.org/10.1177/0146167201272002

Robinson, G. E., Fernald, R. D., & Clayton, D. F. (2008). Genes and social behavior. *Science, 322*(5903), 896–900. https://doi.org/10.1126/science.1159277

Robitzsch, A. (2020) Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education, 5*, 589965. https://doi.org/10.3389/feduc.2020.589965

Ruef, A. M., & Levenson, R. W. (2007). Continuous measurement of emotion: The affect rating dial. J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 286–297). New York: Oxford University Press.

Rumbold, J. L., Dunn, J. G. H., & Olusoga, P. (2022) Examining the predictive validity of the Grit Scale-Short (Grit-S) using domain-general and domain-specific approaches with student-athletes. *Frontiers in Psychology, 13*, 837321. https://doi.org/10.3389/fpsyg.2022.837321

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin*, *125*(1), 3–30. https://doi.org/10.1037/0033-2909.125.1.3

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*(6), 995–1008. https://doi.org/10.1037/0022-3514.91.6.995

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology*, *59*, 419–450. https://doi.org/10.1146/annurev.psych.59.103006.093716

Saha, K., Bayraktaroglu, A. E., Campbell, A. T., Chawla, N. V., De Choudhury, M., D'Mello, S. K., Dey, A. K., Gao, G., Gregg, J. M., Jagannath, K., Mark, G., Martinez, G. J., Mattingly, S. M., Moskal, E., Sirigiri, A., Striegel, A., & Yoo, D. W. (2019, May). *Social media as a passive sensor in longitudinal studies of human behavior and wellbeing*. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland.

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science, 2*(2), 107–114. https://doi.org/10.1177/2515245919838781

Schafer, M., & Schiller, D. (2018). Navigating social space. *Neuron, 100*(2), 476–489. https://doi.org/10.1016/j.neuron.2018.10.006

Schilbach, L. (2014). On the relationship of *online* and *offline* social cognition. *Frontiers in Human Neuroscience, 8*, 278. https://doi.org/10.3389/fnhum.2014.00278

Schmid, P. C., & Amodio, D. M. (2021). Effects of high and low power on the visual encoding of faces. *Social Neuroscience, 16*(3), 293–306. https://doi.org/10.1080/17470919.2021.1906745

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievet, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology, 31*(1), 43–53. https://doi.org/10.1016/j.newideapsych.2011.02.007

Schödel, R., & Mehl, M. R. (2025). Mobile sensing methods. In H. T. Reis, T. West, & C. M. Judd (Eds.). *Handbook of research methods in social and personality psychology* (3rd ed; pp. 297-321). Cambridge University Press.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22*(2), 127–160. https://doi.org/10.1016/S1098-2140(01)00133-3

Schwarz, N., Strack, F., Hippler, H.-J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology, 5*(3), 193–212. https://doi.org/10.1002/acp.2350050304

Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy Scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35–37). NFER-NELSON.

Scollon, C., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. In E. Diener (Ed.) *Assessing well-being: The collected works of Ed Diener* (pp. 157–180). Springer. https://doi.org/10.1007/978-90-481-2354-4_8

Seebacher, F., & Krause, J. (2011). Epigenetics of social behavior. *Trends in Ecology & Evolution, 34*(9), 818–830. https://doi.org/10.1016/j.tree.2019.04.017

Shapiro, D., Jamner, L. D., Lane, J. D., Light, K. C., Myrtek, M., Sawada, Y., & Steptoe, A. (1996). Blood pressure publication guidelines. *Psychophysiology, 33*(1), 1–12. https://doi.org/10.1111/j.1469-8986.1996.tb02103.x

Sharifian, F., Schneider, D., Arnau, S., & Wascher, E. (2021). Decoding of cognitive processes involved in the continuous performance task. *International Journal of Psychophysiology, 167*, 57–68. https://doi.org/10.1016/j.ijpsycho.2021.06.012

Shattuck, E. C. (2021). Networks, cultures, and institutions: Toward a social immunology. *Brain, Behavior, and Immunity Health, 18*, 100367. https://doi.org/10.1016/j.bbih.2021.100367

Shelton, J. N. (2000). A reconceptualization of how we study issues of racial prejudice. *Personality and Social Psychology Review*, *4*(4), 374–390. https://doi.org/10.1207/S15327957PSPR0404_6

Sheng, Y., & Sheng, Z. (2012). Is coefficient alpha robust to non-normal data? *Frontiers in Psychology, 3*, 34. https://doi.org/10.3389/fpsyg.2012.00034

Shepard, L.A. (1997), The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–24. https://doi.org/10.1111/j.1745-3992.1997.tb00585.x

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Shoemaker, P. J., Tankard, J. W., & Lasorsa, D. L. (2004). *How to build social science theories*. Sage Publications.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sijtsma, K. (2009). On the use, misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 103–120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika, 86*, 843–860. https://doi.org/10.1007/s11336-021-09789-8

Sijtsma, K., & van der Ark, L. A. (2020). *Measurement models for psychological attributes*. CRC Press. https://doi.org/10.1201/9780429112447

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein. L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment, 31*(4), 557–566. https://doi.org/10.1037/pas0000648

Sirola, A., Nuckols, J., Nyrhinen, J., & Wilska, T. A. (2022). The use of the Dark Web as a COVID-19 information source: A three-country study. *Technology in Society, 70*, 102012. https://doi.org/10.1016/j.techsoc.2022.102012

Slaney, K. L., & Racine, T. P. (2013). What's in a name? Psychology's ever evasive construct. *New Ideas in Psychology, 31*(1), 4–12. https://doi.org/10.1016/j.newideapsych.2011.02.003

Smith, R. H., & Harris, M. J. (2006). Multimethod approaches in social psychology: Between- and within-method replication and multimethod assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 385–400). American Psychological Association. https://doi.org/10.1037/11383-026

Snijders, T. A. B., & Kenny, D. A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships, 6*(4), 471–486. https://doi.org/10.1111/j.1475-6811.1999.tb00204.x

Snyder, H. R., Friedman, N. P., & Hankin, B. L. (2021). Associations between task performance and self-report measures of cognitive control: Shared versus distinct abilities. *Assessment, 28*(4), 1080–1096. https://doi.org/10.1177/1073191120965694

Song, J., Howe, E., Oltmanns, J. R., & Fisher, A. J. (2023). Examining the concurrent and predictive validity of single items in ecological momentary assessments. *Assessment, 30*(5), 1662–1671. https://doi.org/10.1177/10731911221113563

Spalding, L. R., & Hardin, C. D. (1999). Unconscious unease and self-handicapping: Behavioral consequences of individual differences in implicit and explicit self-esteem. *Psychological*

Science, 10(6). 535–539. https://doi.org/10.1111/1467-9280.00202

Spiegel, A. (2011, May 26). Can a test really tell who's a psychopath? *NPR*. https://www.npr.org/2011/05/26/136619689/can-a-test-really-tell-whos-a-psychopath

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Consulting Psychologists Press. https://doi.org/10.1002/9780470479216.corpsy0943

Spörrle, M., & Bekk, M. (2014). Meta-analytic guidelines for evaluating single-item reliabilities of personality instruments. *Assessment, 21*(3), 272–285. https://doi.org/10.1177/1073191113498267

Stachl, D., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences, 117*(30), 17680–17687. https://doi.org/10.1073/pnas.192048411

Steele, C. M., & Aronson, J. (1998). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811. http://dx.doi.org/10.1037/0022-3514.69.5.797

Steele, C. M., Critchlow, B., & Liu, T. J. (1985). Alcohol and social behavior: II. The helpful drunkard. *Journal of Personality and Social Psychology, 48*(1), 35–46. https://doi.org/10.1037/0022-3514.48.1.35

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 667–680. https://doi.org/10.1126/science.103.2684.677

Stevens, S., Hofmann, M., Kiko, S., Mall, A. K., Steil, R., Bohus, M., & Hermann, C. (2010) What determines observer-rated social performance in individuals with social anxiety disorder? *Journal of Anxiety Disorders, 24*(8), 830–836. https://doi.org/10.1016/j.janxdis.2010.06.005

Stone, A. S., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (Eds.) (2000). *The science of self-report: Implications for research and practice*. Erlbaum.

Strube, M. J., & Newman, L. C. (2007). Psychometrics. In J. T. Cacioppo, L. G. Tassinary, & G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 789–811). Cambridge University Press. https://doi.org/10.1017/CBO9780511546396

Suel, E., Polak, J. W., Bennett, J. E., & Ezzati, M. (2019). Measuring social, environmental and health inequalities using deep learning and street imagery. *Scientific Reports, 9*, 6229. https://doi.org/10.1038/s41598-019-42036-w

Sultana, M., Al-Jefri, M., & Lee, J. (2018). Using machine learning and smartphone and smartwatch data to detect emotional states and transitions: Exploratory study. *JMIR Mhealth Uhealth, 8*(9), e17818. https://doi.org/10.2196/17818

Sun, J., Gan, W., Chao, H.-C., Yu, P. S., & Ding, W. (2022). Internet of behaviors: A survey. *arXiv.2211.15588*. https://doi.org/10.48550/arXiv.2211.15588

Suppes, P., Krantz, D., Lute, D., & Tversky, A. (1989). *Foundations of measurement, Vol. 2*. Academic Press.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In D. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, Volume I* (pp. 3–76). Wiley.

Swendsen, J. D., Tennen, H., Carney, M. A., Affleck, G., Willard, A., & Hromi, A. (2000). Mood and alcohol consumption: An experience sampling test of the self-medication hypothesis. *Journal of Abnormal Psychology*, *109*(2), 198–204. https://doi.org/10.1037//0021-843x.109.2.198

Tal, E. (2021). Two myths of representational measurement. *Perspectives on Science*, *29*(6), 701–741. https://doi.org/10.1162/posc_a_00391

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality, 72*(2), 271–322. https://doi.org/10.1111/j.0022-3506.2004.00263.x

Tay, L. (2018, August 13). *Experience sampling method (ESM) ecological momentary assessment (EMA) webinar* [Video]. YouTube. https://youtu.be/Y80JwwMX3ts

Tay, L., & Drasgow, F. (2012). Theoretical, statistical, and substantive issues in the assessment of construct dimensionality: Accounting for the item response process. *Organizational Research Methods*, *15*(3), 363–384. https://doi.org/10.1177/1094428112439709

Tay, L., & Jebb, A. T. (2018). Establishing construct continua in construct validation: The process of continuum specification. *Advances in Methods and Practices in Psychological Science*, *1*(3), 375–388. https://doi.org/10.1177/2515245918775707

Tay, L., & Kuykendall, L. (2017). Why self-reports of happiness and sadness may not necessarily contradict bipolarity: A psychometric review and proposal. *Emotion Review*, *9*(2), 146–154. https://doi.org/10.1177/1754073916637656

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18*(1), 3–46. https://doi.org/10.1177/1094428114553062

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. A. (2022). Conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment. *Advances in Methods and Practices in Psychological Science, 5*(1), Article 25152459211061337. https://doi.org/10.1177/25152459211061337

te Braak, P., van Tienoven, T. P., Minnen, J., & Glorieux, I. (2023). Data quality and recall bias in time-diary research: The effects of prolonged recall periods in self-administered online time-use surveys. *Sociological Methodology, 53*(1), 115–138. https://doi.org/10.1177/00811750221126499

Temple, D. E., & Geisinger, K. F. (1990). Response latency to computer-administered inventory items as an indicator of emotional arousal. *Journal of Personality Assessment, 54*(1–2), 289–297. https://doi.org/10.1207/s15327752jpa5401&2_27

Thapa, S., Tay, L., & Hou, D. (2021). Experience sampling methodology: Conceptual and technological advances for understanding and assessing variability in well-being research. In P. D. Harms, P. L. Perrewé, & C.-H. Chang (Eds.), *Examining and exploring the shifting nature of occupational stress and well-being* (Vol. 19, pp. 137–154). Emerald Publishing Limited. https://doi.org/10.1108/S1479-355520210000019007

Thun, E., Bjorvatn, B., Osland, T., Steen, V. M., Sivertsen, B., Johansen, T., Lilleholt, T. H., Udnes, I., Nordhus, I. H., & Pallesen, S. (2012). An actigraphic validation study of seven morningness-eveningness inventories. *European Psychologist, 17*(3), 222–230. https://doi.org/10.1027/1016-9040/a000097

Tomarken, A. J. (1995). A psychometric perspective on psychophysiological measures. *Psychological Assessment, 7*(3), 387–395. https://doi.org/10.1037/1040-3590.7.3.387

Torres Irribarra, D. (2021). *A pragmatic perspective of measurement*. Springer. https://doi.org/10.1007/978-3-030-74025-2

Trafimow, D. (2016). The attenuation of correlation coefficients: A statistical literacy issue. *Teaching Statistics, 38*(1), 25–28. https://doi.org/10.1111/test.12087

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, *19*(5), 579–599. https://doi.org/10.1177/0959354309341926

Triandis, H. C., & Suh, E. M. (2002). Cultural influences on personality. *Annual Review of Psychology*, *53*, 133–160. https://doi.org/10.1146/annurev.psych.53.100901.135200

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology, 7*, 769. https://doi.org/10.3389/fpsyg.2016.00769

Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science, 23*(6), 466–470. https://doi.org/10.1177/0963721414550706

Turner, C. F., Miller, H. G., & Rogers, S. M. (1997). Survey measurement of sexual behavior: Problems and progress. In J. Bancroft (Ed.), *Researching sexual behavior: Methodological issues* (pp. 37–60). Indiana University Press.

Uebersax, J. S. (2006). *Likert scales: Dispelling the confusion*. http://john-uebersax.com/stat/likert.htm

Uher, J. (2022). Functions of units, scales and quantitative data: Fundamental differences in numerical traceability between sciences. *Quality & Quantity, 56*, 2519–2548. https://doi.org/10.1007/s11135-021-01215-6

Uher, J. (2023). What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Social and Personality Psychology Compass*, e12740. https://doi.org/10.1111/spc3.12740

van Bork, R., Rhemtulla, M., Sijtsma, K., & Borsboom, D. (2024). A causal theory of error scores. *Psychological Methods, 29*(4), 807–826. https://doi.org/10.1037/met0000521

van der Maas, H. L. J., Kan, K.-J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. Seriously. *Journal of Intelligence*, *2*(1), 12–15. https://doi.org/10.3390/jintelligence2010012

van Fraassen, B. C. (1980). *The scientific image*. Clarendon.

van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research, 144*, 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLoS One, 8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

Verbeij, T., Pouwels, J. L., Beyens, I., & Valkenburg, P. M. (2021). The accuracy and validity of self-reported social media use measures among adolescents. *Computers in Human Behavior Reports, 3*, 100090. https://doi.org/10.1016/j.chbr.2021.100090

Vernham, Z., Tapp, J., & Moore, E. (2016). Observer ratings of interpersonal behavior as predictors of aggression and self-harm in a high-security sample of male forensic inpatients. *Journal of Interpersonal Violence, 31*(9), 1597–1617. https://doi.org/10.1177/0886260515569060

Viswanathan, M. (2005). *Measurement error and research design*. Sage Publications.

Vize, C. E., Miller, J. D., & Lynam, D. R. (2021). Examining the conceptual and empirical distinctiveness of Agreeableness and "dark" personality items. *Journal of Personality, 89*(3), 594–612. https://doi.org/10.1111/jopy.12601

Vogels, J., Demberg, V., & Kray, J. (2018). The Index of Cognitive Activity as a measure of cognitive processing load in dual task settings. *Frontiers in Psychology, 9*, 2276. https://doi.org/10.3389/fpsyg.2018.02276

Walentynowicz, M., Schneider, S., & Stone, A. A. (2018). The effects of time frames on self-report. *PLoS ONE, 13*(8), e0201655. https://doi.org/10.1371/journal.pone.0201655

Wallis, J. D. (2018). Decoding cognitive processes from neural ensembles. *Trends in Cognitive Science, 22*(12), 1091–1102. https://doi.org/10.1016/j.tics.2018.09.002

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology*, *39*(5), 456–467. https://doi.org/10.1016/s0022-1031(03)00019-2

Warnes, E. D., Sheridan, S. M., Geske, J., & Warnes, W. A. (2005). A contextual approach to the assessment of social skills: Identifying meaningful behaviors for social competence. *Psychology in the Schools, 42*(2), 173–187. https://doi.org/10.1002/pits.20052

Watson, D., & Clark, L. A. (1991). Self- versus peer ratings of specific emotional traits: Evidence of convergent and discriminant validity. *Journal of Personality and Social Psychology, 60*(6), 927–940. https://doi.org/10.1037/0022-3514.60.6.927

Watson, D., & Walker, L. M. (1996). The long-term stability and predictive validity of trait measures of affect. *Journal of Personality and Social Psychology, 70*(3), 567–577. https://doi.org/10.1037/0022-3514.70.3.567

Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making, 15*(4), 263–290. https://doi.org/10.1002/bdm.414

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion, 17*(2), 267–295. https://doi.org/10.1037/emo0000226

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research, 49*(5), 737–747. https://doi.org/10.1509/jmr.11.0368

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18*(3), 320–334. https://doi.org/10.1037/a0032121

Weir, J. P. (2006). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research, 19*(1), 231–240. https://doi.org/10.1519/15184.1

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS ONE, 11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719S

Wijsen, L. D., Borsboom, D., & Alexandrova, A. (2022). Values in psychometrics. *Perspectives on Psychological Science, 17*(3), 788–804. https://doi.org/10.1177/17456916211014183

Wise, J. (2023, February 21). *How much data is generated every day in 2023?* Earthweb. https://earthweb.com/how-much-data-is-created-every-day/

Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II. A search procedure to locate the greatest lower bound. *Psychometrika, 42*(4), 579–591. https://doi.org/10.1007/BF02295980

Woods, S. A., & Hampson, S. E. (2005). Measuring the Big Five with single items using a bipolar response scale. *European Journal of Personality, 19*(5), 373–390. https://doi.org/10.1002/per.542

Wu, C.-H., & Yao, G. (2007). Examining the relationship between global and domain measures of quality of life by three factor structure models. Social Indicators Research, 84(2), 189–202.

https://www.jstor.org/stable/20734515

Wu, L., Waber, B. N., Aral, S., Brynjolfsson, E., & Pentland, A. (2008, December). *Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an IT configuration task*. In Proceedings of the International Conference on Information Systems, Paris, France.

Xu, H., Zhang, N., & Zhou, L. (2020). Validity concerns in research using organic data. *Journal of Management, 46*(7), 1257–1274. https://doi.org/10.1177/0149206319862027

Yarkoni, T. (2020). Implicit realism impedes progress in psychology: Comment on Fried (2020). *Psychological Inquiry, 31*(4), 326–333. https://doi.org/10.1080/1047840X.2020.1853478

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yoo, S. J. (2010). Two types of neutrality: Ambivalence versus indifference and political participation. *Journal of Politics, 72*(1), 163–177. https://doi.org/10.1017/S0022381609990545

YouTube (n.d.). *Product features: Recommended videos*. https://www.youtube.com/howyoutubeworks/product-features/recommendations/

Zarate, J. M. (2023). fMRI signatures of social perception. *Nature Neuroscience, 26*, 1. https://doi.org/10.1038/s41593-022-01248-6

Zeng, B., Wen, H., & Zhang, J. (2020) How does the valence of wording affect features of a scale? The method effects in the Undergraduate Learning Burnout Scale. *Frontiers in Psychology, 11*, 585179. https://doi.org/10.3389/fpsyg.2020.585179

Zeng, J., & Schäfer, M. S. (2021). Conceptualizing "dark platforms". Covid-19-related conspiracy theories on 8kun and Gab. *Digital Journalism, 9*(9), 1321–1343. https://doi.org/10.1080/21670811.2021.1938165

Zhang, X., Zhou, L., & Savalei, V. (2023). Comparing the psychometric properties of a scale across three Likert and three alternative formats: An application to the Rosenberg Self-Esteem Scale. *Educational and Psychological Measurement, 83*(4), 649–683. https://doi.org/10.1177/00131644221111402

Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement, 69*(4), 548–565. https://doi.org/10.1177/0013164408324469

ENDNOTES

1. From Spiegel's (2011) interview of Hare on NPR's *All Things Considered*, which focused on the need to measure psychopathy in order to study it and Hare's ambivalence about the use of measures outside the research context. ↑