Arbitrary Metrics in Psychology

Hart Blanton — *University of North Carolina at Chapel Hill*
James Jaccard — *Florida International University*

Many psychological tests have arbitrary metrics but are appropriate for testing psychological theories. Metric arbitrariness is a concern, however, when researchers wish to draw inferences about the true, absolute standing of a group or individual on the latent psychological dimension being measured. The authors illustrate this in the context of 2 case studies in which psychologists need to develop inventories with nonarbitrary metrics. One example comes from social psychology, where researchers have begun using the Implicit Association Test to provide the lay public with feedback about their "hidden biases" via popular Internet Web pages. The other example comes from clinical psychology, where researchers often wish to evaluate the real-world importance of interventions. As the authors show, both pursuits require researchers to conduct formal research that makes their metrics nonarbitrary by linking test scores to meaningful real-world events.

Keywords: clinical significance, validity, reliability, Implicit Association Test, prejudice

Measurement is a cornerstone of psychological research and practice. Measures of psychological constructs are used to test theories, to develop and evaluate applied intervention programs, and to assist practicing psychologists in making treatment decisions. Test developers and researchers alike are careful to document the reliability and validity of their measures, relying on traditional statistics like test–retest correlations, alpha coefficients, correlations reflecting convergent and discriminant validity, and various other forms of validity coefficients. In the present article, we discuss a facet of measurement that is distinct from reliability and validity but that receives scant attention from applied psychologists: the metric of measures. Further, we analyze the arbitrariness of those metrics. We begin by characterizing the nature of metrics in psychological research, highlighting the arbitrary nature of many of them. We then discuss strategies that can be used to make arbitrary metrics less arbitrary. Two case studies are considered that allow us to frame important issues underlying the use of arbitrary metrics. One case study focuses on a research-sponsored Web page that uses response latencies to diagnose unconscious attitudes and prejudices. The second is the use of arbitrary metrics when determining the effects of interventions in clinical research. We conclude with recommendations for future research focused on arbitrary metrics.[1]

The Nature of Arbitrary Metrics

Most constructs in psychology are hypothetical in character and not directly observable. Psychologists cannot observe an individual's standing on variables such as depression, prejudice, self-esteem, and job satisfaction. Instead, they infer location on such dimensions by observing the individual's behaviors. These behaviors can take many forms, but, more often than not, they are the ratings a person gives to items on psychological inventories.

Although such measures provide only indirect assessments of psychological constructs, the hope is that they provide sufficient information to test psychological theories. When measuring self-esteem using a rating scale, for instance, the researcher assumes that a person providing a rating of 6 on a 1-to-7 scale is conveying information about his or her true self-esteem. When measuring romantic attraction in terms of seating distance, the researcher assumes that couples sitting six feet apart from one another are saying something about their degree of attraction for one another. In each case, the researcher uses an assessment technique that yields an observable score, and this score is said to represent an individual's standing on a theoretical and unobservable psychological dimension. The score quantifies a person's standing on the psychological construct in terms of amount, polarity, degree, or magnitude. The term *metric* refers to the numbers that the observed measures take on when describing individuals' standings on the construct of interest. For the dimension of self-esteem, for instance, the metric might range from the

Hart Blanton, Department of Psychology, University of North Carolina at Chapel Hill; James Jaccard, Department of Psychology, Florida International University.

We thank Danny Axsom, Curtis Hardin, Chet Insko, Andy Karpinsky, Jim Neely, Brett Pelham, Diederik Stapel, Penny Visser, and the graduate students in the social psychology programs at the University of Albany and the University of North Carolina at Chapel Hill for comments and assistance with earlier versions of this article.

Correspondence concerning this article should be addressed to Hart Blanton, who is now at the Department of Psychology, Texas A&M University, College Station, TX 77843-4235. E-mail: hblanton@gmail.com

[1] Issues surrounding arbitrary metrics can be framed from the vantage point of different measurement theories, including theories of representational measurement (Luce, Krantz, Suppes, & Tversky, 1990; Suppes, Krantz, Luce, & Tversky, 1989), Rasch measurement (Rasch, 1980; van der Linden & Hambleton, 1997), and classic psychometric theory (Lord & Novick, 1968). We use the latter approach because of its widespread familiarity to psychologists, but we recognize that the approach is not without its critics (e.g., Kline, 1998; Michell, 2000).

Vol. 61, No. 1, 27–41 DOI: 10.1037/0003-066X.61.1.27

lowest possible rating of 1 to the highest possible rating of 7, with larger numbers presumed to indicate higher self-esteem. For the dimension of romantic attraction, the metric might range from the closest observed seating distance of one foot to the largest observed seating distance of seven feet, with larger numbers indicating lower levels of romantic attraction. In each case, the units should not be misconstrued as the true units of the unobserved psychological dimension. A person's true self-esteem is no more a point on a rating scale than a couple's true romantic attraction is the space between them.

Metrics in psychological research often are arbitrary. We define a metric as arbitrary when it is not known where a given score locates an individual on the underlying psychological dimension or how a one-unit change on the observed score reflects the magnitude of change on the underlying dimension. This definition of metric arbitrariness makes explicit that an individual's observed score on a response metric provides only an indirect assessment of his or her position on the unobserved, hypothetical psychological construct. It is assumed that some response function relates the individual's true score on the latent construct of interest to his or her observed score on the response metric (Lord, 1952; Lord & Novick, 1968; Nunnally, 1978). When a metric is arbitrary, the function describing this relationship and the parameter values of that function are unknown.

Consider as an example a depression inventory that locates people's depression on an observed metric from 0 to 50. Suppose an individual receives a score on this index of 35. Knowing only this number and where it falls on the metric in no way conveys how depressed this individual is. This score may occur for people who meet formal definitions of clinical depression, or it may occur for people who show few noteworthy signs of depression. Until psychologists know what psychological reality surrounds the different scores on this scale, the response metric is arbitrary.

Figure 1 presents a hypothetical example to help clarify the above discussion. In Figure 1, the top line represents the true underlying dimension of marital satisfaction, with a neutral point reflecting neither satisfaction nor dissatisfaction. As one moves away from the neutral point to the left, there are increasing amounts of marital dissatisfaction, and as one moves away from the neutral point to the right, there are increasing amounts of marital satisfaction. The first scale, Scale X, is a six-item agree–disagree scale that yields integer scores that tap into a person's true marital satisfaction. This scale has a metric that ranges from 0 to 6. One can determine how these seven numeric values map onto the true dimension of marital satisfaction by extending the lines bracketing a number up to the true dimension of marital satisfaction. If a person is located on the true dimension anywhere between the two lines (extended upward) for a given numeric category, then the person is assigned the number in that category. For the numerical categories at the two extremes of the scale, there is no outer line limiting the category from the dimensional extremes. This produces category widths that are extended at the extremes, a result that is often observed in psychophysical scaling. Figure 1 shows that on Scale X, Person A receives a score of 0 and Person B receives a score of 3, but that each individual would receive different scores on Scale Z. Although Scale Z's metric also ranges from 0 to 6, the content and structure of its items are such that its central categories tap into a wider range of the true underlying dimension, with larger category widths at each numeric value. Thus, Person A has a score of 2 on Scale Z and Person B has a score of 4.[2]

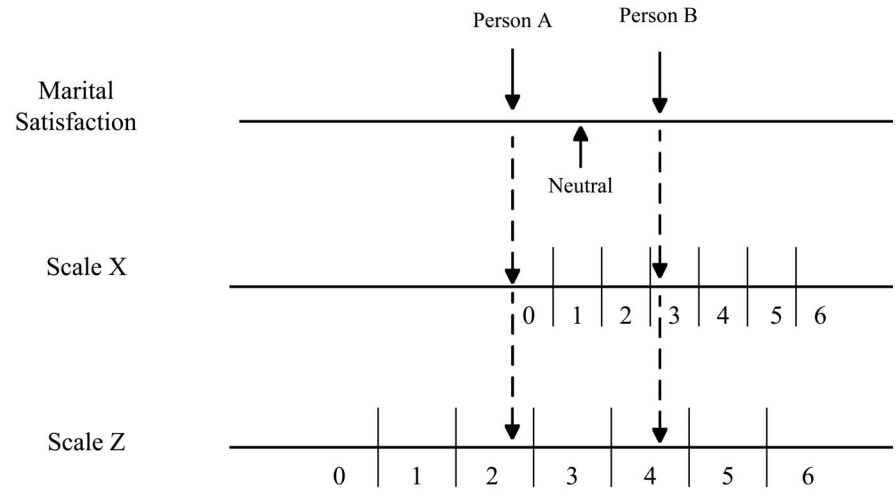
Figure 1 illustrates why one should not try to infer true extremity simply on the basis of the extremity on an observable metric. For instance, the observed score of 0 on Scale X, by and of itself, does not allow one to infer Person A's location on the dimension of marital satisfaction. It would be a mistake to assume that because the value of 0 is the lowest value on the scale, a person with a score of 0 is completely dissatisfied with his or her marriage. It becomes evident that this is not the case when one sees the position of such a person on the true dimension of marital satisfaction in Figure 1. The same is true of Scale Z. A person's observed score on this scale, by and of itself, does not allow one to make a formal inference about the location of the individual on the true underlying dimension. A person who scores 0 on Scale X can score 0, 1, 2, or 3 on Scale Z, and these scores are perfectly consistent with one another when one recognizes that the metrics are arbitrary. It is also important not to infer that the midpoints of Scale X and Scale Z (the score of 3) map onto the midpoint of the true psychological dimension. Indeed, it turns out that the midpoint of Scale X does not map onto the theoretical neutral point, but the midpoint of Scale Z does overlap with it.

One also must be cautious not to make inferences regarding the magnitude of change on a true psychological dimension based simply on the magnitude of observed change on the observed metric. Note in Figure 1 that a change of one unit for Scale X, say from a score of 2 to a score of 3, reflects a different amount of change in true marital satisfaction than a one-unit change for Scale Z (using the midpoint of the category widths as a reference point). With arbitrary metrics, one has no idea how much a one-unit change in the observed metric translates into change on the underlying dimension. It reflects some degree of change, but how much is unknown.

Psychologists must grapple with the dynamics illustrated in Figure 1 because they typically cannot observe the constructs that they want to study. In contrast, physical dimensions often are observable, and this helps in the creation of metrics for them that are not arbitrary. Consider as an example the convention of describing an individual's height in terms of feet and inches. Because height can be observed directly and because people have a great deal of

[2] Both scales are not strictly continuous in that there is coarseness due to the category widths and the collapsing of individuals with different true scores into the same category. This is common for many psychological measures, and researchers typically assume that the coarseness is not problematic.

Figure 1
Two Scales With Arbitrary Metrics



experience using the metric of feet and inches in the real world, people have an appreciation for the meaning of a person's height when expressed in these units. On the basis of experience, for example, we have a good sense of the true height difference between an individual who is 6 feet 1 inch tall and an individual who is 6 feet 7 inches tall. By contrast, when height is expressed in meters, the metric becomes arbitrary for many Americans because they have little sense of the amount of height that a meter represents. They do not know how the metric maps onto the underlying dimension. Psychological metrics can be arbitrary in this same sense. Researchers cannot observe the true psychological quantities that scores represent, and they often have too little experience working with the metrics to know the meaning of any given value. This lack of insight is an unfortunate reality of psychological research. However meaningful psychological constructs may be, the inventories used to measure them often speak to psychologists in an unknown language of arbitrary metrics.

Valid, Reliable, and Arbitrary

Studies exploring the validity of a scale can sometimes help to provide meaning to a metric, but issues of metric arbitrariness are distinct from those of reliability and validity. As an example, both of the scales in Figure 1 could accurately and reliably sort individuals along the true dimension of marital satisfaction, but the links between observed scores and true scores can remain unknown, making their metrics arbitrary. Consider another example. Suppose we tell you that we developed a new measure of height. Unbeknownst to you, this measure is a simple transformation of height as measured in feet. Specifically, it is (feet + 200)(40). Suppose we tell you that an object or entity has a score of 8,400 on the new measure. Knowing nothing else, this score is uninterpretable. Suppose, however, we further tell you that a one-story house has a score of 8,800, that your father has a score of 8,240, and that a newborn infant has a score of 8,060. Associating these external referents with specific scores begins to make this new metric less arbitrary. You begin to gain a sense of how scores on the metric map onto the true underlying dimension. Note that the new measure is perfectly reliable and valid. However, in the absence of links to the external referents, the scores themselves are meaningless.

The example using a new metric for height also makes evident that the issue of arbitrariness is distinct from that of predictive validity. Predictive validity refers to a set of strategies that scientists use to convince skeptics that variability in the observed scores reflects variation on the true underlying dimension. To the extent that a measure predicts phenomena it should predict, one has increased confidence in the validity of the measure. Note that the new (and arbitrary) height measure would have predictive validity in that it would predict phenomena that height is supposed to predict. However, computing a correlation coefficient between a scale and a criterion for the purpose of establishing the validity of that measure is quite a different enterprise from linking meaningful referents to specific scores so as to imbue a metric with meaning.[3] Typically, researchers seek to reduce metric arbitrariness only after a measure has already been accepted as being reasonably reliable and valid.

[3] The term *predictive validity* is used in different ways by researchers, but the most common characterization is the way we have described. The use of prediction to determine if the systematic variance captured by a measure reflects what it is supposed to is not the same as saying that a measure is useful because it predicts important outcomes (e.g., a measure is useful because it predicts the probability of suicide).

Conceptualizing Regions of Dimensions

Setting measurement aside, it can be difficult to gain a sense of different regions on psychological dimensions unless these regions themselves are carefully defined and conceptually elaborated upon. For example, what does it mean to be "high" on the dimension of depression? What does it mean to have a "very negative" evaluation on the dimension of prejudice against African Americans? If a dimension has no inherent metric, then a sense of the meaning of different regions of that dimension (e.g., being low on the dimension) may need to be elaborated. Just as a construct takes on more meaning as it is embedded within a broader nomological network, so will the regions of a dimension take on meaning as they are associated with external events and other variables. As one approaches the issue of arbitrary metrics, it may be necessary to elaborate the meaning of different regions of the conceptual dimensions rather than relying on a more global definition of the construct per se.

Arbitrary Metrics in Psychological Research

For many research purposes, the use of measures with arbitrary metrics is not problematic. This is true when the focus of research is on the study of basic theoretical processes and the desire is to test for the presence or absence of predicted linkages between theoretical variables. Investigators who follow this tradition typically have no particular interest in characterizing a given research participant as being high or low on the underlying dimension or in characterizing changes on a dimension as being small, medium, or large. Rather, the researcher is interested in knowing whether the numerical scores that are assigned to individuals have properties that permit the application of certain statistical methods (e.g., parametric statistics), which, in turn, can be used to determine if scores pattern themselves in ways consistent with known psychological theories. Consider as an example a researcher who believes that similarity leads to attraction. This can be studied in a laboratory context by testing if experimental manipulations of attitude similarity influence seating distance between two individuals (Allgeier & Byrne, 1973). Evidence that proximity scores are ordered across conditions in ways consistent with theory is informative, and it is of little consequence whether any of the experimental groups can aptly be characterized as being either low, medium, or high in attraction or if the group differences are small, medium, or large.

Nonarbitrary Metrics in Psychological Research

However, there are situations in which psychologists want a better understanding of the location of an individual or a group of individuals on a psychological dimension; that is, they desire a metric that is not arbitrary. As the above illustrates, the type of evidence required to confer metric meaning differs from that needed to study basic psychological processes. It is not sufficient simply to test theoretical associations between variables. One must also pursue research that ties specific scores on a metric to specific events that are meaningful in the life of the respondents. Of course, what some scientists or practitioners will view as meaningful events in the life of their respondents may not be considered meaningful by others. Some form of consensus on the part of the scientific or applied community dictates whether an event will be thought of in this way. We consider the role of consensus in more detail later. For now, we simply note that meaning is gained by linking a scale metric to meaningful events that are of applied interest and that this process is not necessarily tied to the goal of testing psychological theories.

Although metric meaning typically is gained by linking scores to external events, this does not prevent researchers at times from trying to infer nonarbitrary meaning on the basis of suspect sources of information. We consider two such strategies in this article. We call the first method *meter reading*. With meter reading, researchers simply use the score on the observed metric to infer location on the underlying dimension. For example, someone at the high end of a metric might be viewed as high on the theoretical dimension and someone at the low end might be considered low on the theoretical dimension. The second strategy is *norming*. In this strategy, raw scores are transformed into standardized scores or percentiles on the basis of normative data and interpretations are imposed on the basis of this new metric. As we show, neither of these approaches is sufficient to generate nonarbitrary metrics and both can placate researchers into believing that metric meaning has been addressed.

In this article, we consider two contexts in which arbitrary metrics are important, one at the individual level and one at the group level. At the individual level, researchers or practitioners often wish to make a psychological diagnosis on the basis of an individual's observed score on a psychological inventory. Common examples of this are when inventories are thought to measure psychological dimensions that have clinical, health, or educational implications. We consider an example from social psychology, where researchers have begun providing the lay public with feedback about their "hidden prejudices" via popular Internet Web pages. The other example focuses on cases in which researchers wish to evaluate the real-world importance of psychological interventions for groups of people. In clinical psychology, for instance, it is not enough simply to determine if scores pattern themselves across conditions in accord with known theories. There also is interest in documenting whether changes in scores indicate meaningful movement along the true theoretical construct of interest and whether the changes have practical and real-world implications.

Measuring Bias in Milliseconds

> Probably a large proportion of people visiting [the IAT Web site] do not consider themselves to be prejudiced and they are taking the tests to learn about these ordinarily hidden associations that in some cases could produce unintended discriminatory behavior. (Anthony Greenwald, as quoted in J. Schwarz, 1998, ¶ 5)

The preceding quote appeared in a 1998 press release announcing the opening of a Web site funded by the National Science Foundation and National Institute of Mental Health (J. Schwarz, 1998). This site (now located at https://implicit.harvard.edu/implicit/) presented the public with an opportunity to take the *Implicit Association Test* (IAT). The IAT is a cognitive task that ostensibly diagnoses implicit attitudinal preferences that people possess but may not fully appreciate. According to the information provided on the site,[4] more than three quarters of the people who have taken a given test have discovered that they possess implicit preferences for Whites over Blacks or for the young over older adults or that they implicitly endorse gender stereotypes about the relative abilities of men and women.

Of course, feedback of this sort can be disconcerting to people who felt, prior to taking the IAT, that they did not possess these preferences or prejudices. The Web site offers advice for these individuals. Visitors who read the answers to the frequently asked questions find that

> There are two reasons why direct (explicit) and indirect (implicit) attitudes may not be the same. The simpler explanation is that a person may be *unwilling* to accurately report some attitude. For example, if a professor asks a student "Do you like soap operas?" a student who is fully aware of spending two hours each day watching soap operas may nevertheless say "no" because of being embarrassed (unwilling) to reveal this fondness. The second explanation for explicit–implicit disagreement is that a person may be *unable* to accurately report an attitude. For example, if asked, "Do you like Turks?" many Germans will respond "yes" because they regard themselves as unprejudiced. However, an IAT may reveal that these same Germans have automatic negative associations toward Turks. (This IAT result has been demonstrated quite clearly in Germany.) Germans who show such a response are unaware of their implicit negativity and are therefore unable to report it explicitly. (https://implicit.harvard.edu/implicit/uk/faqs.html, response to Question 11)

In essence, the public is told that they may have hidden preferences or prejudices and that the IAT can tap into and diagnose these.

The original IAT Web site is commonly used in psychology courses and various sensitivity workshops to teach people about their unknown biases, and two new Web sites have been introduced to further these goals. One was designed to help "fight hate and promote tolerance" (http://www.tolerance.org/hidden_bias/index.html). The site contends that

> We believe the IAT procedure may be useful beyond the research purposes for which it was originally developed. It may be a tool that can jumpstart our thinking about hidden biases: Where do they come from? How do they influence our actions? What can we do about them? (http://www.tolerance.org/hidden_bias/tutorials/02.html, ¶ 3)

The Web site encourages visitors to take the IAT so they may learn what views "may be lingering in your psyche" (http://www.tolerance.org/hidden_bias/index.html, ¶ 5) and it presents an expanded set of tests that are said to reveal hidden racial, age, body-image, and sexual-orientation biases. A more recent site (http://www.understandingprejudice.org/iat/) was developed to supplement an anthology titled *Understanding Prejudice and Discrimination* (Plous, 2002). This site encourages visitors to take the IAT because it can help them "probe unconscious biases" (¶ 2). According to the information reported on these two Web sites, the majority of people taking these new measures have been told that they possess "automatic preferences" that suggest hidden biases.[5]

In the seven years since its unveiling, the original Web site has administered the IAT over a million times (Greenwald, Nosek, & Banaji, 2003), and we presume that the secondary sites have administered it many more. Combined with the additional publicity gained through news coverage on the major TV networks, articles circulated by the Associated Press, and a Discovery Channel program titled "How Biased Are You?" (Sawyer, 2000), the IAT is gaining influence in the public domain. Given this wide dissemination, it is worthwhile to examine more closely the methods researchers have adopted to infer metric meaning with the IAT. We begin by describing the fundamental task used by the IAT to measure preferences. We then consider methods used to infer metric meaning and contrast them with more formal methods for doing so.

The Measurement of Implicit Preference

Different forms of the IAT have been designed to measure a range of psychological constructs (see Greenwald et al., 2003). In our analysis, we focus on the test designed to measure implicit racial preferences, which we refer to as the *race IAT*. The race IAT allegedly measures hidden preferences by determining how quickly a person can classify different experimental stimuli into one of two categories. The experimental stimuli are words or pictures related to categories that are of interest to the researcher. These stimuli are shown, one at a time, on a computer screen, and visitors to the Web site are asked to categorize them by pressing one of two buttons. One button refers to the category *White or pleasant* and the other button refers to the category *Black or unpleasant*.[6] If the stimulus presented on the computer screen is either a picture of a White face or a pleasant word, then the respondent presses a

[4] All quotes taken from the Internet were correct at the time this article was submitted for publication.

[5] All of the sites offer some disclaimer about the potential validity of feedback given. However, the disclaimers are brief and may not have the force needed to cause respondents to discount the detailed feedback they receive (Ross, Lepper, Strack, & Steinmetz, 1977). They each state that the researchers who designed the test and their respective institutions "make no claim for the validity of these suggested interpretations." Nevertheless, visitors are provided with these researchers' interpretations and links to scientific articles said to indicate test validity. They are also told about the impressive credentials of the scientists who designed the test, including their affiliations with some of the most prestigious universities in the world. Many of those who take the test may be swayed by these credentials and, when combined with a limited background in psychological theory and psychometrics, may accept the interpretations given.

[6] For ease of presentation, we focus only on the version of the race IAT that uses the terms *Black* and *White* to represent race and *pleasant* and *unpleasant* to represent valence. Other versions of the test represent race using the terms *African American* and *European American* or the terms *Black American* and *White American*. Valence has also been represented using the terms *positive* and *negative*.

specified key on a computer keyboard. If the stimulus presented is either a picture of a Black face or an unpleasant word, then the respondent presses a different key. The time it takes a person to make the classification is recorded, and this response latency is the fundamental unit of analysis. This task is often referred to as the *compatible judgment* to denote that it should be easy for someone who prefers Whites to Blacks. It is performed by the individual multiple times with a host of stimuli. In a second task, respondents are asked to classify words or pictures in the same fashion, except one of the keys is associated with the category *White or unpleasant* and the other with the category *Black or pleasant*. This task is commonly referred to as the *incompatible judgment* to denote that it should be difficult for someone who prefers Whites to Blacks. Again, the task is performed multiple times across numerous stimuli.

Preference for one group over the other is determined by computing the "IAT effect." The IAT effect is the mean latency for the incompatible trials minus the mean latency for the compatible trials after certain transformations of the measures have been applied. People who perform the compatible judgment faster than the incompatible judgment are said to have an automatic preference for White people, whereas those who perform the incompatible judgment faster than the compatible judgment are said to have an automatic preference for Black people. Visitors to the Web sites also are provided with feedback about the magnitude of any preferences they possess. On the basis of the size of the difference between the two response latencies, respondents are told they possess either a "slight," a "moderate" or a "strong" preference. According to documentation provided by the Web site, 73% of respondents are given feedback indicating that they have at least some degree of preference for White people, with fully 43% being told that they have a "strong automatic preference for White people." These results have been interpreted by some in the research community, in psychology textbooks, and in the popular media as evidence that the majority of people in our society possess a degree of unconscious prejudice against people who are Black.

Meter Reading

The arbitrary nature of many psychological metrics often is readily apparent. When told that someone has a self-esteem score of 8 on a multi-item scale that ranges from 0 to 10, most psychologists would know not to draw any conclusions regarding the person's absolute level of self-esteem. Although an 8 is relatively high on the range of possible scores, this value may not indicate high self-esteem in any meaningful sense. The arbitrary nature of this metric is obvious and the fallacy of interpreting it as such is apparent. At times, however, psychological inventories are designed in ways that their surface features seem deceptively informative. The IAT is an example of this.

Researchers who use the IAT to provide diagnoses rely, in part, on the logic of meter reading. They simply examine a person's IAT score along the range of possible scores and then imbue these values with meaning relative to the score of 0. Positive scores are seen as a preference for Whites, negative scores as a preference for Blacks, and scores of 0 as indicating no preference. Willingness to engage in meter reading with this inventory could arise from two basic misconceptions about the nature of arbitrary metrics. The first is a belief that a metric that is nonarbitrary when used to measure a physical dimension, such as time, is also nonarbitrary when used to measure a totally different psychological dimension, such as prejudice. The second is a belief that the zero point on a bipolar metric is inherently meaningful and maps onto the true zero point on a bipolar construct. We consider both beliefs in turn and show how they do not justify meter reading.

Physical metrics. The IAT measures preferences in milliseconds, a nonarbitrary metric when used to measure the duration of a response latency. Milliseconds provide a direct and familiar assessment of time, and a metric of milliseconds is meaningful as a quantifier of response latencies. This use of a physical metric might seem like an improvement over more traditional attitude-rating scales, because physical metrics allow one to make objective statements about the nature of a response. For example, one can say that a person with a compatible score of 400 milliseconds and an incompatible score of 800 milliseconds is twice as fast on the first judgment as compared with the second. In contrast, one cannot say that an attitude rating of 6 conveys twice as strong an attitude as an attitude rating of 3.

The metric of milliseconds, however, is arbitrary when it is used to measure the magnitude of an attitudinal preference. An attitudinal preference for one group over another is no more an expression of milliseconds than it is an expression on a rating scale. The function describing the relationship between the underlying attitudinal dimension and the response metric is unknown, even if it is assumed that the IAT is a completely valid indicator of attitudes and that units of milliseconds are a nonarbitrary measure of response latency. The arbitrariness of this as an attitude metric is underscored by the fact that IAT researchers (as well as other cognitive psychologists working with response latencies) use log transformations and other scoring algorithms that alter the metric of the response latency per se (Greenwald et al., 2003). It is hard to imagine why one would transform response latencies as measured in milliseconds to get a better estimate of time, but there are many reasons why one might want to do this to obtain a better estimate of an attitude.

Behavioral counts are another example where a metric that is nonarbitrary for one dimension is arbitrary for another dimension. The number of times a person engages in a particular behavior can provide a meaningful description of an event that is of interest in its own right. For example, a researcher may count the number of times that middle school students have tried marijuana in the past week. The measure has a metric that ranges from 0 to a positive integer probably less than 25, and the different numbers are inherently meaningful as reflections of frequency. But if this same measure is used to index a dimension of risk-taking propensity, then one has no sure sense of how

different scores on this metric map onto the underlying dimension. All manner of counts can be meaningful when used to quantify a physical reality but arbitrary when used to quantify psychological constructs. Counts of discrete behaviors (e.g., cigarettes smoked), physical symptoms (e.g., number of migraines per month), and daily experiences (e.g., number of exams per semester) can provide nonarbitrary information about an individual that nonetheless becomes arbitrary when used to index psychological attributes (e.g., smoking attitudes, stress, and self-threat).

In sum, there is no assurance that a metric that is used to measure a physical dimension will be nonarbitrary when it is used to measure a psychological dimension. Thus, the IAT metric should not be considered meaningful simply because it uses a nonarbitrary metric of time. Rather, the meaning of different IAT scores must be established through research that links specific scores to the observable events that are relevant to the underlying psychological dimension of interest. In the case of the race IAT, this means that its metric becomes meaningful to the extent that one knows just how much "relative implicit preference for Whites versus Blacks" is implied by any given IAT score.

Arbitrary zero points. Many of the constructs that psychologists study are bipolar in character, in that the two ends of the theoretical dimension are assumed to be polar opposites. For such concepts, there is a theoretical midpoint that represents a neutral, indifferent, or equivalent characterization. For example, attitudes are often viewed as bipolar constructs ranging from very unfavorable to very favorable, with a true midpoint of neutral. Social comparison researchers often are interested in assessing comparative evaluations to determine the extent to which people feel above average or below average on a dimension that assumes a true midpoint of average. Following these traditions, the IAT is thought to measure a bipolar theoretical construct that reflects the preference for Whites relative to Blacks. The assumed zero point on this theoretical dimension is that of no preference, with deviations on one side reflecting a preference for Whites over Blacks and deviations on the other side reflecting a preference for Blacks over Whites.

Researchers often want to identify the particular number on the observed metric that corresponds to the true neutral point on the underlying bipolar dimension. For rating scales, some researchers assume that it is simply the midpoint of the scale. For example, if a bipolar rating scale ranges from 1 to 7, the number 4 is assumed to map onto the true zero point. This logic is pervasive in research ostensibly showing egocentric preferences for the self relative to others (Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995; Weinstein, 1980) and research documenting judgmental and knowledge overconfidence (see Erev, Wallsten, & Bedescu, 1994, for discussion). Figure 1 reveals why such assumptions often are not warranted. Faith in this assumption also is questioned by research on the cognitive processes that influence rating scales.

When people use rating scales to answer questions, they must first make cognitive representations of the question asked and their response to it, and they then must translate this judgment into a response on the rating scale provided by the investigator (see N. Schwarz, 1999; Tourangeau, Rips, & Rasinski, 2000). Large bodies of literature in psychophysics and psychometrics indicate that processes may be operating that cause the midpoint of a metric to shift away from the true zero point of the underlying dimension. For example, it has been shown that the ratings people give to identical stimuli are influenced by the number and nature of the stimuli that have been rated just prior to it, the number of categories on the rating scale, the anchors that are used to exemplify the endpoints of the rating scales, the adjectives that are used as the scale endpoints, the adverbs (if they are used at all) that describe scale categories (e.g., *slightly*, *moderately*, *very*), the sheer frequency with which stimuli occur in the real world, and category activation processes (e.g., French-Lazovik & Gibson, 1984; Gannon & Ostrom, 1996; Hardin & Birnbaum, 1990; Rotter & Tinkleman, 1970; Skowronski & Carlston, 1989; Wedell & Parducci, 1988, 2000). Such processes can conspire to make the mapping of the true zero point onto the metric complex.

In the case of the IAT, researchers must assume that processes such as the above do not undermine the zero point of their metric: They must assume that a difference of 0 between two arbitrary metrics (i.e., the incompatible reaction time minus the compatible reaction time) corresponds to the true zero point on the attitudinal dimension of interest (e.g., no attitudinal preference for Whites vs. Blacks). No empirical evidence has been presented to support this assertion but, rather, it is assumed to be true by fiat. This might be due to a belief that a correct zero point will be identified by taking the difference of two separate metrics (i.e., the compatible and incompatible metrics). We examine this logic in more depth.

For rating scales, there certainly are cases where one can compute a difference score that probably will map reasonably well onto a true zero of a psychological inventory. Suppose a marketing researcher asks consumers to rate how much they like two different music CDs, A and B, with each rated on a scale that ranges from 1 (*do not like*) to 7 (*like extremely*). When studying preference, one could isolate a rational zero of *no preference* by subtracting the rating for B from the rating for A. The researcher would have a new metric ranging from -6 to $+6$ that represents preference for A over B. It seems likely that the midpoint of this new scale, 0, would correspond to the true theoretical midpoint that represents no preference for A over B.

But it also is possible that one or more of the response biases mentioned above could operate to cause the measured zero point to shift away from the true theoretical midpoint. This can be evaluated empirically. The researcher, for instance, might let participants choose a CD to take home with them. If the computed zero point maps on to the theoretical zero point, then those with positive scores should choose CD A, those with negative scores should choose CD B, and those with a score of 0 should choose CD A about 50% of the time. If this pattern of data is not observed, then this suggests that the location of no CD preference probably is not at the computed zero point.

Despite the reasonableness of differencing in the above example, there are instances in which differencing will not yield a theoretical midpoint. For example, Biernat and Manis (1994) have shown that different referent standards are invoked by respondents when rating perceptions of competence for men versus women. Because of stereotypes that women are not as competent as men for a wide range of tasks, women often are given higher competence ratings than men are for exhibiting the same level of ability. In such cases, a difference score of 0 on observed ratings of competence would not indicate that a man and a woman were viewed by a rater as being equal in competence. It would indicate instead that the woman was perceived as more competent than the man. As an example, Greenwald et al. (2002) measured self-esteem by having respondents rate themselves on six pleasant-meaning words and six unpleasant-meaning words on 7-point scales with anchors 1 = *not at all characteristic of you* and 7 = *extremely characteristic of you*. Greenwald et al. averaged the responses to the positive items and also averaged the responses to the negative items and then subtracted the latter from the former. According to Greenwald et al., this measure had a "rational zero" because it was a difference score (p. 12). Suppose, however, that the positive items are more positive in nature than the negative items are negative. Following Anderson (1981), suppose further that each item has a true scale value that reflects its degree of positivity or negativity on the dimension of self-worth. If the average absolute scale value of the positive items is larger than the average absolute scale value of the negative items, then someone who equally endorses the positive items and the negative items might actually have positive self-regard rather than neutral self-regard. Despite the difference scoring, the zero point obtained by taking a difference of two measures does not map onto the true zero point because the scale values of the two sets of items are not equally polarized.[7]

If a researcher is interested in identifying the measured value corresponding to the true zero, one must move beyond simple meter reading and develop a theory that makes predictions about how data for other variables should pattern themselves as one moves across the dimension of interest and through the true zero point. To isolate where the true zero point occurs on an arbitrary metric, the theory should predict a distinct data pattern for that zero point (as in our CD example). When one consistently observes the data pattern that is predicted for the true zero point at a particular scale point, then one has a basis for interpreting that number as mapping onto the true zero point. (For a more detailed discussion of this strategy, see Anderson, 1981.)

For the IAT, the conditions needed for a zero difference to map directly onto the true theoretical zero point are nontrivial. Both the compatible and the incompatible response latencies are influenced by two implicit attitudes (the implicit attitude for Whites and the implicit attitude for Blacks). They also are influenced by a person's general processing speed. In general, some people are faster than others at reaction-time tasks, and this factor influences their speed on both the compatible and the incompatible judgments, perhaps differentially so. The positive words used in the IAT task also may be more positive in character than the negative words are negative, and the stimuli depicting Blacks may be more prototypical of Blacks than the stimuli depicting Whites are prototypical of Whites. The net result of such factors is that the set of conditions required for an observed zero score to map directly onto a true neutral preference are extensive and quite possibly not met (see the Appendix).

The value on the IAT metric that maps onto the true zero should be established empirically and not embraced as a measurement assumption. To gain empirical perspective on this issue, it is necessary to develop a coherent theory that makes predictions about how data for observable variables pattern themselves on the two sides of the zero point. Much like establishing a preference for CDs, researchers could identify the IAT score that acts as a psychological dividing line between a behavioral preference for Blacks and a behavioral preference for Whites. One could try to identify, for instance, the computed IAT score that differentiates a Black versus White hiring preference, a Black versus White friendship preference, or even a Black versus White romantic preference. If the value that delineates these choices proves to be anything other than zero, then this would suggest that the computed value of zero does not map onto the true zero point representing a lack of preference for Whites versus Blacks. Although studies have investigated the predictive validity of the race IAT with regard to racial attitudes and prejudicial behaviors (e.g., Greenwald, McGhee, & Schwartz, 1998; McConnell & Leibold, 2001), no published study has shown that the zero point used to diagnose attitudinal preferences is the true dividing line between preference for Blacks versus Whites.

In sum, the assumption that the zero point on the IAT measure maps directly onto the true neutral preference for Whites over Blacks is dubious. Before one makes such assertions, research is needed to verify that this is indeed the case. More generally, one must be cautious about assuming that the midpoint of an arbitrary metric maps onto the true midpoint of the underlying construct.

Norming

One reason researchers develop metric meaning is to categorize individuals in terms of magnitude. With the IAT, for instance, researchers categorize attitudinal preferences as either "slight," "moderate," or "strong." Such labels are best developed through discovery of empirical thresholds that indicate noteworthy changes in the occurrence of observable events tied to the phenomenon in question. For example, many scientists might feel comfortable defining someone as having "strong depression" if their depression scores indicate a high risk of a suicide attempt, whereas

[7] Although it is not always so recognized, when researchers reverse score negative items and then sum responses across all items, they are calculating a difference score between the sum of the positive items and the sum of the negative items.

"slight depression" might seem more appropriate for those who only show signs of lethargy.

Another strategy that researchers use to infer meaning is to standardize raw scores and define extremity on the basis of these normed or transformed values. This logic is incorporated into the IAT. The norming procedure used with the IAT, however, is not based on group norms, that is, on data collected on large groups of individuals. We discuss the case of group norms later. For the IAT, the responses of a single individual across multiple trials are used to compute a standard score, which we hereafter refer to as a *d score*. Specifically, the raw IAT score is used in the numerator and the person's own standard deviation across the various IAT trials is used in the denominator. With this convention, the original zero value in the raw scores is preserved when the transformation to a *d* score is applied, but the response metric above and below the zero point is changed. Drawing on Cohen's definition of small, medium, and large effect sizes, researchers tell IAT respondents that they have a "slight preference" if they have a normed IAT score between 0.20 and 0.50, a "moderate preference" if they have a normed IAT score between 0.50 and 0.80, and a "strong preference" if they have a normed IAT score greater than 0.80 (Brian Nosek, personal communication, August 2002).[8]

Transforming to a normed score in this case does little to change the arbitrary nature of this metric. We have no more sense of how much prejudice or automatic preference there is in an IAT score when it is expressed in standardized units than when it is expressed in milliseconds or logged milliseconds. The arbitrariness of the standardized metric could be reduced if the *d* scores were linked empirically to observable expressions of prejudice. For example, if one finds that individuals with normed IAT scores of 0.20 or greater are typified by nontrivial acts of racial discrimination, then this would help to give meaning to such values. To date, however, no empirical studies have pursued this strategy to empirically ground the thresholds used with the race IAT. The majority of people taking this test are being provided with feedback indicating that they have automatic preferences that might suggest a form of hidden prejudice. These diagnoses are based on scores that are arbitrary, even though they are transformed.

As another perspective on this, suppose that a person has a true automatic preference value that is close to the theoretical neutral point. Suppose further that this individual provides consistent responses across IAT trials, such that his or her standard deviation is very small. Transforming the IAT score to a *d* score could yield a value much larger than 0.80 because researchers would be dividing the IAT score by a very small standard deviation. The respondent would be given feedback indicating a strong automatic preference for one ethnic group over the other, even though the individual's true score may be so close to "no preference" that it has no practical consequences whatsoever. In short, one's *d* score is inversely related to one's standard deviation on the IAT tasks (because the standard deviation is in the denominator of the *d* score), but it is unclear why the standard deviation on the race IAT should be a factor that determines feedback about having strong or weak automatic preferences. The situation is further complicated by the fact that researchers have no real sense of the range of the true underlying dimension to which the IAT is sensitive. Perhaps the IAT task and stimuli are sensitive to only a narrow range of true scores centered around the theoretical neutral point of the bipolar dimension. The IAT scores themselves may show considerable variability, but this variability might map onto only small amounts of variability on the true underlying dimension.

Concluding Comments on the IAT

The IAT is being used in the public domain to diagnose hidden biases and prejudices. However, the arbitrary nature of the IAT metric and the fact that diagnoses have not been linked to any observable acts of automatic preference suggest that researchers have no way of gauging the true magnitude of the implicit preference expressed by a given IAT score. The use of the IAT in its present form to assign psychological diagnoses places undue faith in meter reading and norming. Researchers and practitioners should refrain from making such diagnoses until the metric of the IAT can be made less arbitrary and until a compelling empirical case can be made for the diagnostic criteria used. We outline future research that might be used for this purpose shortly. First, we turn to the issue of assessing the clinical significance of a treatment or intervention.

Clinically Meaningful Results and Arbitrary Metrics

Just as one might wish to gain a sense of where an individual stands on an unobserved psychological dimension, one also might wish to gain some sense of how much a given individual or group of individuals has changed on that dimension as a result of an intervention. A concept in clinical psychology that has been gaining attention is that of *clinical significance* or *clinically significant change*. The concept has evolved from treatment studies and focuses on whether an intervention has a meaningful impact on the everyday life of clients (Kazdin, 1999). The idea is that it is not enough to demonstrate statistically significant mean changes on an outcome measure that reflects a psychological construct (e.g., depression, anxiety, or marital satisfaction). Rather, it also must be shown that those changes have meaningful consequences for individuals.

Meaningful Change

Although it has not been framed as such, the pursuit of clinical significance can be viewed, in part, as an attempt to make arbitrary metrics less arbitrary. For example, when

[8] We have recently learned that the diagnostic criteria for the Web site changed in 2003, on the basis of work done on the new IAT scoring procedure (Greenwald et al., 2003). Specifically, "Values of the new measure that are used as minima for the labels correspond approximately to Cohen *d* values of 0.3 ('slight preference'), 0.7 ('moderate preference') and 1.3 ('strong preference')" (Anthony Greenwald, personal communication, November 2005). No matter how minima for these categories are set, both the new values and the original values are arbitrary.

statistically significant changes on a metric of marital satisfaction are obtained, clinicians have no idea how much true change in marital satisfaction has occurred, nor does the clinician have a sense of the ramifications of that change for couples. If research can tie the metric for marital satisfaction to observable marital experiences, then the metric becomes less arbitrary and one can begin to appreciate what it means to shift from, say, a score of 4 to a score of 5. The emphasis on clinical significance is a direct result of dissatisfaction with the arbitrariness of the metrics of many clinical measures (Sechrest, McKnight, & McKnight, 1996).

Not surprisingly, controversy exists about how one should define clinical significance. Kazdin (1999) noted that many researchers define clinical significance in terms of symptoms and symptom reduction. However, he also pointed out that there are other dimensions on which clinical significance can be defined, such as meeting role demands, functioning in everyday life, and improvement in the quality of one's life. Furthermore, these criteria can be invoked for either the client, the significant others who interact with the client (e.g., a spouse, a parent), or even society at large. It is not our purpose to consider the complex issues in defining clinical significance for different disorders in different clinical settings (see Kazdin, 1999, for a cogent discussion of these issues). Rather, we develop the implications of the use of arbitrary metrics in clinical research when considered in the context of clinical significance.

Norming

Although some clinicians engage in meter reading, research on clinical significance is dominated by an emphasis on norming. This takes varying forms, but it almost always involves calibrating the observed score on a scale to values for some reference group. A common perception is that a metric that is arbitrary somehow becomes meaningful once it is normed. Although it is true that such standardization can convey important and useful information about relative standing, standardization alone cannot convey someone's absolute standing on a psychological dimension of interest, nor does it necessarily calibrate a measure to meaningful external events. To illustrate, we discuss two common norming strategies, one focused on a single group and another focused on multiple groups.

Single-group norms. This approach to defining clinical significance uses standardized z scores to describe how much an individual has changed relative to the mean and standard deviation of some reference group. For example, whereas before a treatment program, a person might be 2.5 standard deviations above the mean for the general population, after the treatment, the person is only 1.5 standard deviations above that mean, representing a change in z-score units of $2.5 - 1.5 = 1.0$ unit in the direction of "normalcy." In this framework, the focus is either on whether the individual has crossed a particular threshold value (e.g., a treatment moves a person below a z score of 2.0) or on the amount of z-score change (e.g., a treatment leads to a decrease of 1.0 z-score unit). Although this latter method of norming alters the units of change, it does nothing to make the metric of change less arbitrary. It also does not speak to the issue of clinical significance.

To illustrate, suppose that one measures the weight change of a group of people undergoing an obesity reduction program. If standard scores truly imbue meaning to arbitrary metrics, then one should be able to gain a sense of actual weight reduction if the reduction is characterized strictly in standard score units, without recourse to the nonarbitrary metric of pounds and ounces. If one is told that an obese person had weight reduction corresponding to a standard score of 1.0, one has no sense of how much weight the person actually has lost. The person may have lost 10 pounds or maybe 50 pounds. The difference in the reality of these two scenarios is great. Because the actual number of pounds lost is not known, the clinical significance of this loss is certainly not known either (e.g., how this loss affects the person's overall health). To gain perspectives on the clinical significance of this 1.0 z-score unit change in weight, one must either (a) convert the normed units back into nonarbitrary units, such as pounds (for which data about clinical significance might exist), or (b) determine how this amount of weight loss in z-score units affects the daily functioning of the individual (thereby linking the z score to externally defined events that are indicative of clinically significant change). Either of these strategies will make the normed metric more meaningful and potentially allow one to gain a sense of the clinical significance of this person's change in weight.[9]

Now return to the former case in which a person shows a change of 1.0 standard score unit toward the mean value of a reference population on a measure of a psychological dimension. Suppose that the referent population for calculating z scores and defining clinically significant change is a normal population, such as a representative sample of the U.S. population.[10] A cutoff score is defined at 2 standard deviations from the mean of the normal population in the direction of the psychopathology. If an individual reliably shifts from being above this cutoff value to being below the cutoff value, then clinically significant change is said to have occurred (see Jacobson, Roberts, Berns, & McGlinchey, 1999, for a discussion of this and other cutoff strategies based on group norms). This approach sidesteps the spirit of clinical significance in that no data are presented to show that those below the cutoff value behave differently or in improved ways relative to those above the cutoff value. No attempt is made to make the z scores less arbitrary by tying them to meaningful, real-world events. One still has no idea whether the treatment has had any meaningful or practical impact on the everyday

[9] One should not infer from our discussion that standard scores, in general, are not useful for determining how extreme a person's score is relative to the scores of other people. They do convey this information, and this can be useful. But a standard score says little about the location of a person on a psychological dimension in an absolute sense or of its behavioral implications.

[10] In practice, the reference groups are often just convenience samples or are somewhat ad hoc in character.

functioning of the individual being treated. Instead, the focus of the analysis is entirely within the arbitrary metric of the outcome variable.

Examining scores in terms of standard deviation units is simply a rescaling of the metric and does not make the metric any less arbitrary. There is no sense of how much the underlying psychological construct has changed when someone's standard score of 2.2 is reduced to a standard score of 1.8, nor is it known if there are any implications of that change for the individual being treated. To quote Kazdin (1999):

> The question for any measure or index of clinical significance is the extent to which the measure in fact reflects a change that does have an impact on the individual's functioning in everyday life or a change that makes a difference. . . . Stated another way, clinical significance is not being measured because researchers call the measures clinically significant or adopt them for convention. . . . Measures of clinical significance require supporting evidence to establish that they actually do reflect important, practical, worthwhile, and genuine changes in functioning in everyday life. (p. 336)

Multiple-group norms. In addition to approaches that define cutoff points on the basis of a single reference group (e.g., a normal population), there also are approaches that focus on multiple reference groups. For example, a researcher may identify two groups of individuals, those who are "dysfunctional" and those who are normally functioning or "functional." Jacobson et al. (1999) described three approaches that have been used to define cutoffs for clinical significance in this context: (a) Use a cutoff that is 2 standard deviations from the mean of the dysfunctional group (in the direction of functionality), (b) use a cutoff that is 2 standard deviations from the mean of the normal or functional group (in the direction of dysfunctionality), and (c) use a cutoff that defines the score where a client is statistically more likely to be defined as being in the functional group as opposed to the dysfunctional group. We have already noted problems with the first two strategies, namely, that they do not relate score changes to actual improvement in the client's life or everyday activities. By contrast, the third strategy has potential for linking the cutoff value to meaningful events if membership in one group or the other is viewed as a proxy for the occurrence of clinically meaningful external events. To the extent that this is the case, then the cutoff score also will reflect these differences and the score will then be empirically tied to real-world events. For a more detailed discussion of multigroup norming approaches, see Kendall, Marrs-Garcia, Nath, and Sheldrick (1999).

Clinical Significance and Nonarbitrary Metrics

Researchers who wish to address the issue of clinical significance are served well by making their outcome measures less arbitrary, but the concept of clinical significance extends beyond that of metric meaning. To illustrate, suppose that the outcome variable in question has a nonarbitrary metric, such as the number of migraine headaches that the person experiences in a month. If one is told that a treatment reduces an individual's score on this metric from 10 to 4, then one has a sense of what this means. It means that on at least six fewer occasions per month, the person is able to escape the debilitating consequences that follow from a migraine headache. This change, most likely, will improve the quality of life for this individual. But suppose the change was from 10 to 9 headaches. Does this change impact the quality of life of the individual, or is it too small? Even though the metric of the outcome is nonarbitrary in terms of the number of headaches, one does not know how variations in this metric map onto other important psychological dimensions, such as the individual's overall quality of life. This example shows that clinical significance is an issue that transcends the arbitrariness of a metric. For clinical significance, the real-world impact of an intervention must be documented, whether the metric is arbitrary or nonarbitrary. Nevertheless, a by-product of work addressing the question of clinical significance is that it helps reduce the arbitrariness of metrics.

Arbitrary Metrics and Indices of Effect Size in Clinical Research

The concept of clinical significance and our discussion of arbitrary metrics underscores the somewhat vacuous nature of psychology's recent emphasis on standardized effect size indices (Hevey & McGee, 1998; Matthey, 1998; Thompson, 2002). Effect size estimation characterizes the relative difference of two or more groups on an outcome measure (e.g., a treatment group vs. a control group). The impression given by advocates of standardized indices of effect size is that they somehow make the practical significance of an effect more evident. The difficulty with such logic can be seen by considering the standardized effect size of a two-group treatment effect using Cohen's *d*. The unstandardized measure of effect size (using sample notation) is simply the difference in means for the two groups, namely $M_1 - M_2$. Cohen's *d* is the mean difference divided by the pooled standard deviation of the two groups, $(M_1 - M_2)/s$. A small effect size, according to Cohen (1988), is one whose *d* value is 0.20, and this standard is often invoked when interpreting effect sizes. Note that the only difference between the two indices of effect size is that Cohen's *d* divides the mean difference by *s*, whereas the unstandardized index does not. Is it really the case that the simple act of dividing a mean difference by a standard deviation reveals the practical, real-world implications of that difference? Of course not. Dividing by the pooled standard deviation does nothing more than rescale the unstandardized difference onto another metric. The new metric is just as arbitrary as the original.

Judging the practical significance of an effect size requires a researcher to link empirically the metric of the effect size to the practical and tangible costs and benefits that can be observed. Whether one chooses to do this when the metric is in standardized or unstandardized form is a minor point. What is important is making the empirical links between the effect-size metric and observable criteria that have clinical significance, thereby rendering the effect-

size index less arbitrary. Standardized effect-size indices, although potentially useful, can be counterproductive in that investigators can be lulled into a false sense of metric meaningfulness. Suppose, for instance, that a treatment versus control manipulation yields a Cohen's *d* of 0.80. This effect might be deemed important because it represents what Cohen calls a large effect. It is entirely possible, however, that the intervention in question had minimal effects in terms of changes on the true underlying dimension of interest and that these changes had no consequential effects in the lives of the individuals being treated.

To illustrate, suppose that female assistant professors at a university are paid an average of $50,000 and male assistant professors are paid an average of $50,001, but the pooled standard deviation is just 1.0 (indicating almost no variability in salaries). The standardized effect size for the mean difference is (50,001 − 50,000)/1.0 = 1.0. A researcher applying Cohen's criteria would call this a large effect size. But the absurdity of inferring meaningful sex discrimination becomes evident when one reverts to the original metric. The metric of dollars is nonarbitrary and with it one can see that the spending and lifestyle implications of the gender gap in pay are trivial.

These comments should not be taken as an indictment of effect-size estimation strategies. The basic premise of magnitude estimation frameworks is relevant for many areas of research. Our point is simply that meaning does not appear when one indexes the magnitude of a treatment effect in terms of standardized units. One must also link these units to tangible, observable events in the real world. This is the crux of the often-stated distinction between statistical significance and practical significance. Unfortunately, it is rare to find carefully reasoned accounts of when a particular effect size reflects an effect with practical significance. Instead, researchers typically default to the criteria suggested by Cohen (1988).[11]

Conclusions

In the present article, we have suggested a criterion other than the traditional ones of reliability and validity that test developers and applied researchers often should take into account. This focuses on the arbitrariness of the metric of a test or scale. As noted at the outset of this article, matters of metric arbitrariness are of minor consequence for theory testing and theory development, but they can be important for applied work when one is trying to diagnose an individual's absolute standing on a dimension or when one wishes to gain a sense of the magnitude and importance of change. What does it mean when one obtains an 8-unit change on a depression scale that ranges from 0 to 50? What are the clinical implications of a score of 20? To reduce arbitrariness, test developers should build a strong empirical base that links specific test scores to meaningful events and that defines cutoff or threshold values that imply significantly heightened risks or benefits. This task can present many challenges. Unlike physical concepts (e.g., height), psychological constructs often have no agreed-upon referents that convey absolute standing on the underlying dimension of interest. Nevertheless, one can begin to impart meaning to a metric by linking different scores to observable referents or markers that are thought to vary on the dimension of interest. We illustrated how this can be done in our example with the new measure of height. This metric was made less arbitrary when observable events (e.g., objects differing in height) were associated with different points on the metric.

Difficulties that can arise during this phase of test development are that scientists or practitioners may disagree on what markers are important and where these markers fall along the underlying dimension. The goal should still be to seek consensus so that some perspective on metric meaning can be gained. Psychologists who wish to reduce the arbitrariness of a measure should therefore (a) identify the relevant events they view as meaningful, (b) make a case for the importance of those events and the positioning of those events on the underlying psychological dimension in an absolute sense, (c) build consensus among members of the scientific or applied community about such positioning, (d) conduct the necessary research to link test scores to those events in such a way as to render the metric of the test meaningful, and (e) make a case and build consensus for the threshold values used to make diagnostic statements.

These guidelines can imbue a metric with meaning, but it is important to realize that just as the validity of a scale is contextually bound, so too is the meaning of its metric. Technically, validity and reliability are not properties of scales; rather, they are properties of data (Messick, 1995). Thus, the extent to which a set of measures is valid is dependent not only on the scale or instrument used to generate observations but also on the particular set of individuals on which the observations are made, the time at which the data are collected, and the setting in which the data are collected. These same factors put boundaries on the meaning of a metric. Researchers should consider this limitation whenever they seek to generalize inferences about a metric to new research or applied contexts. When necessary, they will have to conduct generalizability studies to delimit the populations and conditions to which a metric's meaning extends.

Two Case Studies

In terms of our two case studies, we believe that it is questionable to use the IAT to provide the public with feedback about their unconscious and hidden stereotypes and prejudices. Even though a scale based on milliseconds is nonarbitrary when used to assess the physical dimension of time, the IAT uses a metric that is arbitrary when it is used to assess such unobservable dimensions as automatic preference. For this reason, one cannot say with any confidence that the zero point of the IAT maps onto the true neutral point of such preferences, nor can one determine how observed deviations from this zero point translate into degrees of true preference. With no research linking the

[11] For a lucid discussion of the limitation of Cohen's criteria, see Lenth (2001).

diagnostic thresholds for the IAT to observable actions that might be related to such preferences, the approach taken on IAT Web sites amounts to little more than meter reading. The IAT has dubious justification as a diagnostic instrument, and we question whether any individual should ever be provided with the kind of feedback given daily to visitors of the IAT Web sites.

In terms of our analysis of clinical significance, the major point we make is that the strategy of forming group norms does not necessarily make an arbitrary measure less arbitrary. Standardization simply rescales one arbitrary metric into another. It is only when a (standardized or unstandardized) metric is tied to clinically important outcomes that the meaning of different scores emerges. This is true for effect-size indices as well, as these scores also must be grounded to external events to become nonarbitrary.

The Value and Challenges of Making Metrics Less Arbitrary

A metric, once made meaningful, can be used to provide perspectives about such things as the magnitude of change that occurs on an underlying dimension as a function of an intervention. Evidence that an intervention causes movement along a scale that has nonarbitrary meaning can reveal the real-world consequences of this change. This assumes, of course, that the mapping of the metric onto external events has not changed as a function of the intervention, but there is always this possibility. For example, interventions could increase concerns for socially desirable responding, alter interpretations of scale anchors, or influence the interpretation of questions being asked. Researchers who address these possibilities can make more confident statements regarding metric meaningfulness and clinical significance.

It can be difficult and time consuming to conduct the research needed to make a metric less arbitrary. Fortunately, the issue of metric arbitrariness is irrelevant for many research goals, so not all researchers must tackle this issue. If one simply wishes to test if variables pattern themselves in ways predicted by a theory, then there usually will be no need to conduct studies to reduce the arbitrariness of the metric. However, there are applied situations in which researchers need to address the issue if they are going to fulfill their research goals. Tying metrics to meaningful, real-word events provides a viable means of making metrics less arbitrary, but there will always be some guesswork involved. No new methodology is going to expose psychological constructs to the naked eye. Best estimates of where people stand on psychological dimensions are always that, estimates. Nevertheless, awareness of this limitation is of value to the psychologist. A researcher who appreciates the gap between a psychological metric and a psychological reality knows to look past a person's score and search for something meaningful.

REFERENCES

Ackerman, P. L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. *Intelligence, 10,* 101–139.

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102,* 3–27.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology, 68,* 804–825.

Allgeier, A. R., & Byrne, D. (1973). Attraction toward the opposite sex as a determinant of physical proximity. *Journal of Social Psychology, 90,* 213–219.

Anderson, N. (1981). *Methods of information integration.* New York: Academic Press.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology, 66,* 5–20.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (in press). Decoding the Implicit Association Test: Implications of conceptual and observed differences scores for criterion prediction. *Journal of Experimental Social Psychology.*

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519–527.

French-Lazovik, G., & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement, 8,* 49–57.

Gannon, K. M., & Ostrom, T. M. (1996). How meaning is giving to rating scales: The effects of response language on category activation. *Journal of Experimental Social Psychology, 32,* 337–360.

Greenwald, A. G., Banaji, M., Rudnam, L., Farnham, S., Nosek, B. A., & Mellott, D. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109,* 3–25.

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,* 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197–216.

Hardin, C., & Birnbaum, M. H. (1990). Malleability of "ratio" judgments of occupational prestige. *American Journal of Psychology, 103,* 1–20.

Hevey, D., & McGee, H. M. (1998). The effect size statistic: Useful in health outcomes research? *Journal of Health Psychology, 3,* 163–170.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67,* 300–307.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 332–339.

Kendall, P. C., Mars-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 285–299.

Kline, P. (1998). *The new psychometrics: Science, psychology, and measurement*. New York: Routledge.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *American Statistician, 55,* 187–193.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph* (Whole No. 7).

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

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement: Vol. 3. Representation, axiomatization, and invariance.* San Diego, CA: Academic Press.

Matthey, S. (1998). $p < .05$—But is it clinically significant?: Practical examples for clinicians. *Behaviour Change, 15,* 140–146.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37,* 435–442.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology, 10,* 639–667.
Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
Plous, S. (2002). *Understanding prejudice and discrimination*. New York: McGraw-Hill.
Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago Press.
Ross, L. D., Lepper, M. R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology, 35,* 817–829.
Rotter, G. S., & Tinkleman, V. (1970). Anchor effects in the development of behavior rating scales. *Educational and Psychological Measurement, 30,* 311–318.
Sawyer, F. (Host). (2000, March 20). *How biased are you?* [Television broadcast]. Silver Spring, MD: Discovery Channel.
Schwarz, J. (1998). 33,000 Web tests show unconscious roots of racism, ageism [Press release]. Retrieved November 11, 2001, from http://www.washington.edu/newsroom/news/1998archive/10–98archive/k100898b.html
Schwarz, N. (1999). Survey methods. In D. T. Gilbert & S. T. Fiske (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 143–179). New York: McGraw-Hill.
Sechrest, L., McKnight, P., & McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. *American Psychologist, 51,* 1065–1071.
Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105,* 131–142.
Suppes, P., Krantz, D. H., Luce, R. D., & Tversky, A. (1989). *Foundations of measurement: Vol. 2. Geometrical, threshold, and probabilistic representations.* San Diego, CA: Academic Press.
Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80,* 64–71.
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey responses.* Cambridge, England: Cambridge University Press.
van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.
Wedell, D. H., & Parducci, A. (1988). The category effect in social judgment: Experimental ratings of happiness. *Journal of Personality and Social Psychology, 55,* 341–356.
Wedell, D. H., & Parducci, A. (2000). Social comparison: Lessons from basic research on judgment. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (pp. 223–252). Dordrecht, the Netherlands: Kluwer Academic.
Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39,* 806–820.

Appendix
Removing IAT Method Confounds by Differencing

This appendix describes the conditions needed for simple differencing to remove a method confound from the Implicit Association Test (IAT). We focus on one known method confound, general processing speed.

In the race IAT, it is reasonable to assume that an individual's true relative preference for Whites over Blacks (RP) impacts both the response latency on the compatible task (CRL) and the response latency on the incompatible task (IRL). This can be expressed more formally as

$$\mathrm{CRL} = \alpha_1 + \beta_1 \mathrm{RP} + \varepsilon_1$$

$$\mathrm{IRL} = \alpha_2 + \beta_2 \mathrm{RP} + \varepsilon_2$$

The latent relative preference has been framed in IAT research as a difference in attitudes, in this case, the attitude towards Whites (A_W) minus the attitude towards Blacks (A_B).

$$\mathrm{RP} = A_W - A_B$$

Substituting, we obtain

$$\mathrm{CRL} = \alpha_1 + \beta_1(A_W - A_B) + \varepsilon_1$$

$$\mathrm{IRL} = \alpha_2 + \beta_2(A_W - A_B) + \varepsilon_2$$

Now suppose that each response latency is also influenced by some other factor that represents method variance. The one we consider here is general processing speed, or the tendency for some individuals to be able to respond more quickly than others across all response latency tasks. If we add a factor for method variance to each equation, we obtain

$$\mathrm{CRL} = \alpha_1 + \beta_1(A_W - A_B) + \beta_3 M + \varepsilon_1$$

$$\mathrm{IRL} = \alpha_2 + \beta_2(A_W - A_B) + \beta_4 M + \varepsilon_2$$

If we compute the difference score and rearrange terms, we obtain

$$\mathrm{CRL} - \mathrm{IRL} = (\alpha_1 - \alpha_2) + (\beta_3 - \beta_4)M + (\beta_1 - \beta_2)(A_W - A_B) + (\varepsilon_1 - \varepsilon_2) \quad \text{(A1)}$$

$$= (\alpha_1 - \alpha_2) + (\beta_3 - \beta_4)M + (\beta_1 - \beta_2)A_W - (\beta_1 - \beta_2)A_B + (\varepsilon_1 - \varepsilon_2) \quad \text{(A2)}$$

It can be seen from this that the effects of a method artifact are removed only if its influence on the two latencies is the same, namely, $\beta_3 = \beta_4$.

For the case of general processing speed, research suggests that this may not be the case and such simple differencing will not remove processing speed confounds. One source of doubt derives from the literature on task performance and task difficulty. As a general principle, individual differences in skill (e.g., general processing speed) will only manifest themselves for tasks that are moderate to high in difficulty. This is because a very easy task will be solved by all and a very difficult task will be solved by none. Thus, neither extremely easy nor extremely difficult tasks will be influenced by individual differences in skill levels. For tasks falling in the middle range of difficulty, however, increased skill will improve performance, with increased skill yielding greater benefits for tasks that are moderately hard as opposed to tasks that are moderately easy (Ackerman, 1986, 1987). The theory sur-

rounding the IAT supposes that incompatible judgments typically will be harder for people than compatible judgments will be. This difference is thought to occur because most people are exposed more often to cultural images that are consistent with compatible judgments as compared with incompatible judgments (Greenwald et al., 1998). If this view is correct, then the literature on task performance suggests that the incompatible judgments will be influenced by general processing speed to a greater degree than the compatible judgments will be.[A1]

Equation A2 also makes evident that in addition to the assumption of equal influence of method artifacts, one also must assume equal intercepts, $\alpha_1 = \alpha_2$, and that the impact of the attitude towards Whites on the two tasks is the same as the influence of the attitudes toward Blacks (because $\beta_1 - \beta_2$ is the estimated coefficient for both terms). It is beyond the scope of this article to address these assumptions in detail. However, it is fair to say that these assumptions require a nontrivial leap of faith, given the lack of empirical research to support them.

[A1] Greenwald et al. (2003) described a new scoring algorithm that supposedly eliminates the bias of differential impact of processing speed. However, the basic premises of the scoring algorithm are open to debate, and there are reasons to believe that the algorithm is not optimal (see Blanton, Jaccard, Gonzales, & Christie, in press, for discussion of this issue).