JOURNAL OF PERSONALITY ASSESSMENT, *85*(2), 112–124

SPECIAL SERIES: Understanding Construct Validity

Conceptual Complexity and Construct Validity

Robert E. McGrath
School of Psychology
Fairleigh Dickinson University

Despite a century of methodological and conceptual advances in the technology of psychosocial measurement, poor correspondence between indicators and the constructs they are intended to represent remains a limiting factor to the accumulation of scientific knowledge. Longstanding conventions in measurement may contribute to the failure to develop optimal criteria. These conventions include the focus on complex over simple constructs and the use of multi-item measures of disparate content to represent those constructs. Several arguments suggest that such a measurement model compromises the potential for developing measures that accurately reflect psychosocial phenomena. The article concludes with some preliminary suggestions concerning an alternative model that may address this construct validity problem more effectively.

> Clarity in the large comes from clarity in the medium scale; clarity in the medium scale comes from clarity in the small. Clarity always comes with difficulty.
>
> —Tukey, 1969, p. 88

One of the basic requirements of science is accurate measurement. Despite a century of effort devoted to improving methods for the creation and evaluation of measurement devices, psychologists generally agree that many if not most of the scales commonly used for the observation of psychosocial events and states provide at best a rough reflection of the constructs they are intended to represent. The concern raised by the lack of congruence between the measure and the measured is reflected in a substantial research literature devoted to the issues of construct validity and the criterion problem.[1] It is reasonable to hypothesize that the continuing lack of correspondence between psychosocial measures and constructs is an important obstacle to progress in the accumulation of scientific knowledge.

In this article, I explore the proposition that the conceptual complexity of the constructs psychologists choose to measure and the scales they use to measure them has played an important role in the failure to develop more accurate measurement systems. The *conceptual complexity* of a construct refers to the degree to which the construct hierarchically encompasses conceptually distinct subconstructs. A construct such as depression carries with it a broad set of implications about more specific components of behavioral, emotional, cognitive, and physiological functioning, many of which may or may not be true in any individual's case. The same could be said of constructs such as job satisfaction and intellectual ability. A common approach to the measurement of complex variables is complex scales that aggregate items relevant to these conceptually distinct subconstructs. The case to be presented is that both types of complexity compromise the potential for accurate and precise characterization of psychosocial phenomena.

This article consists of four parts. The first section deals with the distinction between the predictive and representational purposes of measures. I will attempt to make the case that representational accuracy is primary for the advancement of empirical knowledge. The second section has to do with whether it is theoretically possible to measure conceptually complex constructs with sufficient accuracy. This is followed by a discussion of the practical obstacles resulting from the use of conceptually complex scales. I conclude the article with a discussion of possible strategies for addressing the problem of conceptual complexity.

PREDICTION AND REPRESENTATION

There are two purposes served by psychosocial indicators (Cone, 1995). They can be used to predict status on another variable. This is useful when the target variable is only di-

[1]The term *criterion problem* has also been used to refer to the practical obstacles to accurate measurement (Austin & Villanova, 1992; Wiggins, 1973).

rectly measurable in the future (e.g., future job performance) or the past (e.g., whether the person was "insane" at the time of an offense), or when the variable cannot be directly accessed (e.g., the prediction of diagnosis in the absence of biographical data). A score generated in the context of prediction is interesting for what it says about an external referent. Using the terminology of correlation familiar to psychologists, a predictor is optimal when it covaries strongly with variables it is intended to predict.

A measure can also be used as a representation of a construct. This occurs when the measurement is primarily intended to reflect an individual's location on the construct that ostensibly underlies the measure. A score generated in the context of representation is interesting for what it says about an intrinsic referent. Returning to correlational terms, a representation is optimal when its covariations with optimal representations of other constructs provide a reasonable estimate of parametric values. Cronbach and Meehl (1955) introduced the term *construct validity* in relation to the representational accuracy of scales.

It is worth noting that the concept of construct validity assumes that constructs have a character independent of their measurement. This assumption is consistent with the philosophical perspective often referred to as critical realism (Collier, 1994; Manicas & Secord, 1983). The central tenet of realism is the existence of an objective reality to—in the case of psychology—psychosocial constructs, even if the capacity to measure objective status on those constructs is flawed. Critical realism is not the only perspective available on the independent existence of constructs. Positivists, for example, would argue against the scientific admissibility of a reality beyond that which is consensually observed. This school had its greatest influence on psychological measurement in the form of operationism (Friedman, 1991), which defined the target of a measurement solely in terms of the observable measurement process. Despite the potential for alternative viewpoints, critical realism represents the dominant view among social scientists. It provides the context for this discussion, and I treat the representational accuracy or construct validity of a test as a meaningful concept.

Reading research on test validation might lead one to assume that the predictive purposes of tests are more important than their representational purposes. The bulk of the literature dedicated to demonstrating scale validity has focused on scale relationships with expected correlates. In his classic text on assessment, Wiggins (1973) even asserted that "personality assessment has the quite applied aim of generating predictions about certain aspects of behavior that will contribute to decisions concerning the disposition or treatment of individuals" (p. 6), a statement that overlooks the descriptive and model-building aspects of personality research.

Several factors may contribute to the tendency among psychologists to focus on scales as correlates rather than as representations. Prediction is an important goal in applied settings; the lingering influence of operationism in the discipline might play a role as well. The demonstration of criterion-related validity is also considered an important contribution to the evaluation of construct validity (Cronbach & Meehl, 1955; Messick, 1995), one with the advantage over others in that it can be reduced to a relatively clear-cut statistical strategy.

Whatever the reason for the emphasis on prediction, the representational effectiveness of a test is in fact far more important to a scale's scientific value. Science is largely a process of developing and testing models. Model building in turn requires accurate representations of placement on the constructs included in the model. Any time a scale is used as a criterion or dependent variable, or for descriptive purposes, or for the estimation of parameters, the representational accuracy of the measurement is paramount because some rough equivalence between placement on the variable and placement on the construct is required.[2] If the meaning of the scores on a scale is unclear, then the accuracy of any inferences about constructs made on the basis of that scale is in doubt. Scales that correspond poorly with the constructs they are intended to represent cannot provide the basis for clear answers to empirical questions. In the next section, I evaluate whether it is possible to develop accurate representations of conceptually complex constructs.

CONCEPTUAL COMPLEXITY IN CONSTRUCTS

Trait theorists have long recognized that constructs used to characterize an individual can be organized hierarchically in terms of their degree of complexity or abstraction from particulars (Eysenck, 1947; Hampson, John, & Goldberg, 1986; Paunonen, 1998). At one end are very broad constructs such as the domains of the Five-Factor Model (Goldberg, 1993) that encompass a large array of more specific attributes. At a lower level of generality are these more narrowly defined general qualities such as the cross-situational tendency to experience anxiety. Still further down in terms of complexity are enduring styles of reacting to specific stimulus situations such as the tendency to experience anxiety in a certain way in response to certain events. Constructs at the top of this dimension are complex and integrative. The further down one moves on the dimension, the more elemental and specific constructs become.

[2]Subtle changes in context can turn what looks like a predictive study into a model-building study. For example, hierarchical regression is often used to determine the practical value of an additional predictor. However, hierarchical analysis can also be used for model building when the goal is to determine whether a particular construct adds to the understanding of another. Even though the scale is likely to be called a predictor in this context, its purpose is to serve as a representation of some construct.

Problems With Complex Constructs

The emphasis on complex constructs actually may not be optimal either in the context of prediction or the context of representation. Studies that have compared scales reflecting constructs at different levels of complexity consistently find that prediction is enhanced by using a larger number of more specific personality variables rather than a smaller number of more global ones (e.g., Mershon & Gorsuch, 1988; Paunonen, 1998). In this section, I focus on the more conceptual question of whether representational accuracy is also complicated by the use of global constructs rather than more specific ones. There are several reasons to suspect this is the case.

Complex constructs as social constructions. The construct of personal hopelessness can be understood quite well in terms of a single dimension reflecting the degree to which an individual has negative expectations for the future. It is logically possible to partition the construct further into apprehension about the future and negativity about the potential for change, but self-observation and reports from clinical patients has suggested that the experience of hopelessness is an elemental one. This sense of cohesiveness to the experience suggests that different observers would experience hopelessness similarly, even if they come from different cultural backgrounds.

In contrast, the concept of extraversion is specifically intended to refer to certain covariations among more elemental psychosocial phenomena. A person who experiences great pleasure in social situations is also likely to engage in behaviors leading to social contact and to be perceived by others as outgoing. The problem is that these states do not always cooccur, so the characterization of a person's status in relation to extraversion becomes subjective. Because of personal circumstances, "John" is very frequently involved in social situations and is perceived by others as outgoing, but has no more than a normal enjoyment of those situations. Is John extraverted?

Complicating the matter is the covariation of these three attributes associated with extraversion with many others, so that different observers may weigh subordinate constructs differently. Several authors (e.g., Block, 1995) have pointed out the marked differences across test developers in assumptions about the composition of extraversion. The Revised NEO Personality Inventory (NEO–PI–R; Costa & McCrae, 1992) bases one sixth of its extraversion score on items reflecting warmth, but is warmth even a component of extraversion as opposed to a correlate?

There is no objective answer to either question because extraversion exists only as a social construction (Gergen, 1985), a label used to summarize a loosely bounded set of observed regularities. Under these circumstances, one would expect different cultures and even different observers within the same culture to generate very different definitions of a construct.

Clinical diagnosis (American Psychiatric Association, 2000) represents a particularly important class of social constructs. In a small set of cases, schizophrenia being perhaps the best example (Prescott & Gottesman, 1993), there is good reason to suspect that a disease state existed prior to its observation, and the objective bounds of that disease state will ultimately be defined. In most cases, though, diagnoses reflect a set of covariations between primary and secondary symptoms that were consensually agreed on as the basis for classification. The result is a diagnostic system with an inordinately high rate of comorbidity, heterogeneous classes, and a set of "garbage can" diagnoses for use when no other category applies (Beutler & Malik, 2002; Cooper, 2004).

In fact, many if not most psychiatric phenomena are probably better understood in terms of a single presenting complaint and placement on a set of symptom dimensions potentially affected by that presenting complaint. The disparity between the diagnostic nomenclature and actual psychiatric phenomena is largely ignored, and extensive research is conducted to understand the psychosocial and treatment implications of the existing diagnostic categories, making the diagnostic system one of the best examples of the continuing influence of operationism in psychosocial science (Acton, 1998). To quote Mahrer (1999) who wrote on a related topic, "psychotherapy is a pseudoscience of nonexisting unrealities, measured with rigorous precision" (p. 1150).

The existence of many complex constructs as flexible verbal summaries also has important implications in terms of the potential for building quantitative models of psychosocial phenomena. Michell (2000, 2001) questioned whether there is any evidentiary basis for assuming that psychological constructs can be mapped to ordered numeric values, an argument that would seem particularly relevant to ratings of mental states. If complex constructs do not have an inherent character, then there is no "true" value for parameters such as the correlation between extraversion and depression.

This does not necessarily mean that sample correlations between measures of these two constructs will vary wildly because different measures of extraversion or of job satisfaction usually overlap in their contents. It does have implications for progress in psychosocial theory development. One characteristic of a mature science is the ability to derive point estimates from theory that can be compared to empirical findings. Such comparisons offer strong tests of a theory. In the absence of theory-based point estimates, theory corroboration is reduced to tests of null hypotheses that offer very weak evidence for the validity of a theory (Meehl, 1990).

Excessive faith in cross-modal consistency. The tendency to conceptualize psychosocial phenomena in terms of summative constructs also tends to reinforce excessive faith in cross-modal consistency. The term *mode* is used here to refer to a broad domain of psychosocial activity such as behaviors, self-perceptions, perceptions by others, or physiological functioning. Many complex constructs such as de-

pression, extraversion, or schizophrenia encompass subconstructs from more than one mode of functioning. The tendency to speak in terms of level of extraversion as a comprehensive statement about an individual implies a fair degree of coherence (depending on the researcher's model of extraversion) in personal experience, behavior, perception by others, and/or physiological reactivity. Based on the assumption of cross-modal consistency, Campbell and Fiske (1959) even proposed that finding a scale failed to correlate well with measures of the same construct involving other modes raised questions about the scale's validity.

There are several reasons to expect some degree of cross-modal consistency. The fact that people can observe their own behavior and the effect of their behavior on the social environment, and that events involving the cognitive-affective systems of the brain precede or co-occur with certain physiological and behavioral events, ensures that to some degree, there will be convergence among self-perceptions, perceptions by others, behavioral tendencies, and physiological states. The use of constructs that assume tight integration in all these domains goes well beyond the evidence for cross-modal consistency, however. Physiological measures are poorly related to self-report measures (e.g., Edelmann & Baker, 2002); interpretations based on self-report data relate poorly to clinicians' perceptions (e.g., Ehrenworth & Archer, 1985); parent, teacher, and child reports of a child's areas of difficulty can show an unimpressive rate of agreement (Achenbach, Dumenci, & Rescorla, 2002).

Seeming inconsistencies across modes may actually be an important source of information that has been neglected in the assessment literature (e.g., Borman, White, & Dorsey, 1995; Sayer et al., 1993). It can be important to know whether an employee's supervisors and peers differ in terms of their work evaluations or whether a job candidate believes she or he is much brighter than performance testing suggests. The man who perceives himself as depressed while others perceive him as angry and demanding is different in important ways from the woman who is perceived by both self and others to be sad and guilt ridden. These inconsistencies across modes of measurement are more likely to represent worthwhile clues about a person than evidence of measurement error.

The issue of cross-modal inconsistency also raises concerns about the overreliance on self-report measures as representations of multimodal constructs (Kagan, 1988). The use of self-report alone to represent a complex construct with implications for several modes of functioning is simply inadequate science unless the research is only concerned with the experiential components of that construct.

The position presented here contrasts with the common assumption that a self-report measure can sufficiently represent a complex construct such as anxiety, which involves interpersonal and physiological as well as experiential modes. Rejecting this assumption effectively mitigates long-standing concerns about the validity of self-reports as representations of objective reality (e.g., Meehl, 1945, 1995). If anxiety is perceived as a socially useful label to capture demonstrable correlations among self-report, behavior, and physiological state, then it is clear that a self-report measure provides an insufficient basis for portraying one's standing in terms of severity (or presence/absence) of anxiety. Self-reported anxiety can only provide a representation of the experience of anxiety, although it can also serve as a predictor of anxiety components in other modes.

Inadequate specification in psychosocial models. Even within a mode of measurement, the assumption of consistency across subcomponents interferes with the development of accurate models. The complex construct extraversion bears little resemblance to complex constructs in sciences that have achieved a higher level of representational accuracy. For example, physicists are able to measure physical characteristics very accurately and so have developed a mathematically precise description of how the components of an atom relate to each other. This does not mean they have lost interest in the behavior of atoms as a whole. However, the concept of the atom has evolved into a well-specified set of relationships among subcomponents that provides a basis for understanding why the atom as a whole behaves in the ways it does. Similarly, because currency flows can be measured with a fair degree of accuracy, the relatively young science of econometrics has been able to generate very precise models of the manner in which increased productivity in one sector of the economy affects productivity in others.

In contrast, the specification of extraversion remains largely a laundry list of subconstructs with relatively little known about how one component affects another. The assumption of strong covariation among the subcomponents of extraversion may at least contribute to the lack of research examining relationships among them more closely.

Reasons for Complex Constructs

If measures of more complex constructs are less effective predictors than scales based on more specific constructs (Mershon & Gorsuch, 1988; Paunonen, 1998), and if complex constructs cannot be accurately measured, one must wonder why psychologists continue to focus on complex constructs rather than more specific ones. There are several factors that play a role.

One is the social or pragmatic value of indexing. Economic decision-making is often based on indexes such as the Dow Jones industrial average or the index of leading economic indicators because these provide a global indication of position within a set of related economic variables. Similarly, psychologists respond to various mandates that require categorizing individuals for practical purposes. The designation "mentally retarded" has implications concerning eligibility for social services, as does "insane" for certain forensic

decisions. In fact, many if not most scales in psychology can be reformulated as indexes that exist only for practical purposes. What psychologists sometimes forget, though, is that an index has no inherent structure. The difference between psychological and economic indexes is the degree to which those indexes are perceived as the elemental components of measurement versus a socially useful aggregate of more elemental components.

A related practical rationale is the need to reduce information load for purposes of comprehending information and communicating it to others. Mershon and Gorsuch (1988) hypothesized that many theories suggest five to eight higher order factors can account for much of personality simply because that is the number of variables humans can effectively manipulate at one time. The belief that a person can be characterized in terms of a few global concepts such as depression, intelligence, or responsibility is a seductive one, as it is both easier to grasp and easier to communicate to others than a large set of more specific constructs.

A third factor may be a natural predisposition to organize the world according to prototypes rather than boundary conditions (Rosch, 1973). The prototype of the extravert is a loud, active, sweet, fun-loving person, and there are people who demonstrate all of these characteristics simultaneously, even though some of these characteristics are more likely definitional aspects of the extravert, whereas others are key correlates. This prototype influences how people judge themselves and others. It also influences the development of formal models of extraversion as well as multi-item scales that evaluate extraversion in terms of distance from the prototype (Broughton, 1990). However, the ability to define a prototype does not substitute for the scientific process of defining an objective definition of necessary and sufficient conditions for placement on a construct.

CONCEPTUAL COMPLEXITY AND MULTI-ITEM SCALES

Even if it is possible to provide an objectively defensible model of a complex construct, precise measurement of such constructs may not be possible. Representational accuracy requires a measurement device that reflects the full array of subconstructs the test developer assumes to be relevant. Burisch (1984b) offered one solution to measuring complex constructs that deserves mention because it is sometimes adopted in practice. Burisch experimented with single-item measures of depression that involved, for example, responding on an unanchored 9-point scale ranging from 1 (*often depressed, moody, self-conscious*) to 9 (*contented, self-assured, poised*). Burisch found this type of measure can demonstrate adequate criterion-related validity, sometimes even matching that of much longer measures. However, the construct validity of single-item measures of complex constructs is likely to be questionable. The item leaves it to the respondent to decide what depression is, which symptoms should be considered most salient to making a judgment of intensity, and whether the respondent evaluates level of depression in relation to others' experiences or the respondent's experience of other emotional states. The respondent's interpretation of the item becomes increasingly ambiguous as the construct underlying the item becomes more complex (Hogan & Roberts, 1996).

A much more popular strategy is the multi-item scale involving administration of one or several items representing each of some or all of the [sub]constructs the test developer believes to be subsumed by the construct and combining scores across items. The rationale for the multi-item scale can be traced to the principle of aggregation (Epstein, 1983; Rushton, Brainerd, & Pressley, 1983), which suggests that measurement error is inevitable, but its impact can be reduced by averaging multiple observations. Although it is impossible to disagree with the importance of aggregation as a methodological strategy, the multi-item scale is a relatively unusual instance of the principle. When multiple observers rate the same event or a force is applied to an object multiple times to determine whether the effect is consistent, it is done under the assumption that the repeated observations are redundant, that is, that they are equivalent in what the measurement is intended to represent. This assumption may ultimately prove not to hold in some cases, but the assumption is still central to the aggregation of results. In contrast, multi-item scales regularly sum items that are clearly distinct in their target constructs, although those targets are subsumed by a construct of higher complexity.

Notice that evidence of covariation among items is not the same thing as evidence of conceptual redundancy. Items representing correlated constructs can easily generate a scale that meets statistical criteria for unidimensionality or reliability, particularly if the number of items is fairly large (Schmitt, 1996). Redundancy between items also does not mandate large correlations because redundant items can correlate only moderately if they vary substantially in difficulty. The determination of redundancy, as is true of construct validity in general, is a conceptual one that cannot be reduced to statistical criteria.

Multi-Item Scales As Predictors

Despite Burisch's (1984b) finding, it can be demonstrated mathematically that multi-item scales are likely to be better predictors of criteria on average than single items (e.g., Ghiselli, 1964; Nunnally & Bernstein, 1994). It is sometimes suggested that this superiority is based on the relationship between validity and reliability (e.g., Rushton et al., 1983). The argument goes as follows. Because the square root of reliability defines the upper bound of criterion-related validity and because one way to increase reliability is by increasing the number of items, it follows that a multi-item scale is potentially more valid than a single item.

It is important to note this explanation only applies to the maximum possible validity of a scale, not to its actual valid-

ity. The latter is more accurately modeled as a function of the individual item validities. According to Ghiselli (1964), the correlation between multi-item predictor X and criterion Y equals

$$r_{XY} = \frac{\sum_{i=1}^{k} r_{iY}}{\sqrt{k + k(k-1)\bar{r}_{ii'}}}, \quad (1)$$

where r_{iY} represents the correlation between item i and the criterion, k is the number of items, and $\bar{r}_{ii'}$ is the mean correlation between all pairs of items. Two aspects of this formula are worth mentioning. First, other factors being equal, the criterion-related validity of a multi-item scale declines as the mean correlation between items increases, even though reliability should increase as a function of the mean interitem correlation. Criterion-related validity is actually enhanced by aggregating items that correlate zero or even negatively with each other so long as they correlate positively with the criterion. Second, the process of aggregation reduces the impact of poor predictors on the criterion-related validity of the scale as a whole.

To demonstrate this second point, consider a set of seven items with r_{iY} values equally spaced from .10 to .70 and a mean interitem correlation of .30. The correlation between the aggregated scale and the criterion would equal

$$\frac{.10 + .20 + .30 + .40 + .50 + .60 + .70}{\sqrt{7 + 7(6).30}} = .63.$$

The criterion-related validity of the scale is substantially greater than what would have been expected from a review of the individual item validities.[3]

Although Equation 1 tends to assure reasonable criterion-related validity coefficients, it also has negative implications for the discriminant validity of multi-item scales (Campbell & Fiske, 1959). Depression and paranoia are two constructs likely to be considered conceptually quite distinct. Among the items typically included in a multi-item scale of depression would be some reflecting the respondent's interpersonal perceptions, whereas a paranoia measure is likely to include items having to do with the respondent's sensitivity to criticism by others. These items should correlate, and according to Equation 1, it is possible as a result that the two scales will correlate quite well. The same characteristic that enhances the predictive power of multi-item scales can also produce a nontrivial correlation when such a finding is surprising or even undesirable (e.g., Crowley & Merrell, 2000; Fried-Buchalter, 1992; Neuberg, West, Judice, & Thompson, 1997). However, discriminant validity has more to do with representational accuracy than with criterion-related validity, and consistent with the bias noted earlier in favor of the latter, most studies of scale validity do not even present discriminant validity coefficients. Inflated relationships between multi-item scales of disparate constructs may be an important contributor to the "crud factor," the name Meehl (1990) used for the proposition that "everything correlates to some extent with everything else" (p. 204). It was Meehl's presumption that this phenomenon is inherent to psychosocial phenomena, but the problem may at least be exacerbated by the poor discriminant validity of multi-item scales.

Still another implication of Equation 1 is that the criterion-related validity of multi-item scales levels off after a surprisingly small number of items (see also Burisch, 1984a; Paunonen, 1998; Paunonen & Jackson, 1985; Taylor, Ptacek, Carithers, Griffin, & Coyne, 1972). There is even evidence that for any particular criterion, a shorter and more focused measure will be more valid than the full scale from which it was drawn (e.g., Ashton, 1998; Ashton, Jackson, Paunonen, Helmes, & Rothstein, 1995; Burisch, 1997), even though subscales should on average demonstrate less reliability.

This is not to say there are no situations in which a measure of substantial length might still be considered more useful for predictive purposes. If the 50-item Psychopathic Deviate scale from the Minnesota Multiphasic Personality Inventory–2 (MMPI; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) is considered valuable specifically because it can simultaneously screen for antisocial tendencies, amoral tendencies, egocentric tendencies, anger and impulsive wishes, sensation seeking, the tendency to experience generalized dissatisfaction, and family/authority conflicts, then a large set of items tapping various constructs is clearly called for. If one is going to administer a single measure to tap the entire spectrum of personality or psychopathology, one is better off with a thorough one. The down side of this predictive strategy is that for any one criterion, there is probably a subset of items on the scale that predicts the criterion better than the scale as a whole, and the interpretation of an elevated score on the total scale is very much complicated by the variety of possible explanations. If the purpose of the scale is instead to predict a specific criterion, maximum criterion-related validity is often reached after as few as three to four good items (Burisch, 1997).

Multi-Item Scales As Representations

The bias toward criterion-related validity as an evidentiary basis for construct validity leads test users to assume the larger correlations associated with multi-item scales means they are better scales overall. There is in fact no rationale for assuming that what makes a good predictor makes a good

[3] Another intriguing feature of the formula is that, for the typical length of multi-item scales in psychology (10 to 30 items) and what would seem to be a reasonable mean correlation between individual items and any one behavioral criterion (.20, or about 5% of criterion variance), the correlation of the entire scale and the criterion will reach a maximum in the range .30 to .40. Although not particularly relevant to this discussion, this finding has interesting implications for understanding why there seems to be a ceiling to the criterion-related validity of self-report scales as predictors of behavioral events (see Mischel, 1968).

representation. In fact, the opposite can be true. The aggregation of redundant items, such as multiple measurements of the same event, can reduce the proportion of total variability comprised of measurement error without compromising the accuracy of representation. In contrast, aggregates of items that are distinct in content, as is true of most multi-item scales, have several features that limit their potential as representations.

Issues of scale format. The term *scale format* is intended to encompass such features as the number and content of response alternatives, the wording of the items, the order in which items are presented, the number of items, and so forth. A great deal is known about what constitutes good and bad practice in item and scale formatting. For example, Simpson (1944) demonstrated many years ago that vague markers of frequency such as *frequently* or *occasionally* should be avoided because they are interpreted very differently by different respondents. There is an extensive literature that has demonstrated that responses to individual items can be influenced by supposedly extraneous factors such as item order, but that these effects can be minimized through careful scale development (Sudman, Bradburn, & Schwarz, 1996). Yet there is remarkably little evidence that standardized scale developers are expected to consider this literature when dealing with issues of scale format. Test manuals often provide extensive information about the validity of a scale as a whole but almost no information about the justification for the formatting decisions that were made in the process of test development or even how the mix of item contents was determined.

Neglect of format issues is not specific to aggregates of nonredundant items, nor is it a necessary feature of multi-item scale development. However, the emphasis on prediction over representation surely reinforces the practice. Equation 1 dictates that reasonably well-developed multi-item scales will correlate with relevant criteria adequately, mitigating the need for further attention to detail in the context of prediction. The standard for a scale as a representation is higher because deficiencies in the formatting of the measure compromise its accuracy as a reflection of the person's status relative to the construct.

Issues of isomorphism. Certain isomorphic relationships should exist both within and between scales before a scale can be considered an accurate representation of a construct. First, the use of a scale as a representation implies rough equivalence on the construct among individuals with equal scores. There are over 100 million different combinations of responses that could result in a score of 15 on a simple 30-item, true–false scale. Although many of these combinations are unlikely to occur in practice, it is still realistic to expect that a score of 15 can indicate any of thousands of possible combinations. The assumption of equivalence is therefore only likely to be valid if the items are redundant in content. In the context of prediction, finding that the scale covaries monotonically with important criteria is usually accepted as sufficient evidence of meaningful ordering of scores on the scale; this is not equivalent to demonstrating that equal scores represent the same location on the target construct within the tolerances of measurement error.

Second, alternate measures of the same construct should converge. There are many cases in which this happens when there is a widely held, well-articulated model describing the population of subordinate constructs that are central to the target construct. For example, there is general agreement on the constructs subsumed by depression including suicidality, helplessness, and so forth. It is not surprising then to find that self-report measures of depression can correlate better than .70 with each other (e.g., McGrath & Ratliff, 1993).

For many if not most complex psychosocial constructs, there is no generally accepted model of the relevant subdomains (and in the previous section, I suggested that in many cases there never can be). As a result, scales that ostensibly measure the same construct can show little convergence. For example, the characteristics of emotional distress have been well documented in the clinical literature. Using a sample of 149 college students administered both the NEO–PI and the Interpersonal Style Inventory (ISI; Lorr, 1986), I found the NEO–PI–R Neuroticism scale correlated –.78 with the ISI Stability scale. In contrast, the correlation between the NEO–PI–R and ISI Conscientiousness scales was only .35. MMPI and Millon (1977) Clinical Multiaxial Inventory measures associated with Compulsive Personality Disorder have consistently proven to be negatively correlated (Greene, 2000). Recognizing the problem this lack of isomorphism creates for measures as representations, statisticians interested in building descriptive models have recommended collecting multiple measures of each construct and generating a factor score that better represents the latent construct (e.g., Schumaker & Lomax, 1996). However, the weighting of different subdomains in the estimation of factor scores will still vary depending on the contents of the scales chosen by the researcher.

Finally, when the scale is not developed in a manner consistent with a well-specified construct, it is often unclear whether research findings reveal something about the construct or something about the scale. This is a particularly nettlesome problem when a multi-item scale behaves in an unexpected manner, correlating with the wrong variables or not correlating with the right ones. Many depression scales include items that are conceptually more closely related to anxiety than they are to depression (Gotlib & Cane, 1989). These anxiety items should improve the strength of relationships of depression measures with reasonable criteria because most of the important correlates of depression also correlate with level of anxiety. At the same time, the anxiety items compromise the representational value of the measure by reducing its specificity, which in turn can produce misleading results in the development of real-world models. For example, the corroboration of an attributional model of de-

pression by correlating a measure of depression with a measure of attributional style is compromised if the depression measure is loaded with items that more accurately reflect anxiety (McGrath & Ratliff, 1993).

Similarly, recent evidence has suggested that on average, newer antidepressants only result in a 2-point improvement on the Hamilton (1967) Depression Rating Scale when compared to placebo (Kirsch, Moore, Scoboria, & Nicholls, 2002). When one considers the 21-item version of the Hamilton includes 3 insomnia items, 4 items more closely related to anxiety than to depression, and 3 somatic items having to do with vegetative symptoms, it is not even clear whether these very popular drugs have a specific effect on depression at all.

The inclusion of items that do not correlate with each other or even covary negatively, although potentially enhancing criterion-related validity, complicates the task of interpretation further. For example, Chisholm, Crowther, and Ben-Porath (1997) found an unexpected positive relationship between global level of improvement in psychotherapy and the Psychopathic Deviate scale of the MMPI. Further analyses revealed that the relationship resulted from a positive relationship between improvement and items on the scale that tap feelings of alienation, which are relatively independent of other items on the scale.

Examples involving other constructs besides psychopathology are available as well. Judge, Erez, Bono, and Thoresen (2002) found consistent evidence of extreme overlap between measures of neuroticism, locus of control, and self-esteem. It is possible that these constructs truly overlap to the degree suggested by the results. It could also be the case that the use of multi-item scales with poor discriminant validity for related constructs exaggerates the degree of overlap. When Black Americans demonstrate significantly higher scores than White Americans on a measure of schizophrenia, or scores on an anxiety measure change an average of ½ *SD* in 1 week, the lack of isomorphism between the scale and the construct muddies the waters in terms of whether these findings represent insight or artifact.

Issues of specificity. Multi-item scales can also be accused of modeling psychosocial phenomena in overly broad terms. One example has to do with the evaluation of test–retest reliability. The MMPI Psychopathic Deviate scale combines items reflecting very labile constructs (e.g., dissatisfaction) with others that should be quite stable (e.g., problems with authority). This practice can be criticized in several ways. First, because the expected time frame for change differs across subcomponents, there is no way to identify a reasonable lag to serve as the basis for a test–retest reliability analysis. Second, interesting information about the natural history of the subconstructs is collapsed into a single test–retest reliability coefficient. Third, reasonable changes in subconstructs over time are treated as error in measurement rather than as a worthwhile target of study in its own right. In contrast, because the measurement of individual economic indicators is considered strongly representational, shifts in such indicators can be interpreted as worthwhile information about their natural volatility rather than as error. Similar comments could be made about changes in dependent variable scores in response to psychological intervention: Focusing on overall change can ignore interesting and important differences in the effectiveness of treatment across subdomains represented in the scale.

The impact of imprecision in scales representing complex constructs is also evident in clinical description. For example, the Five-Factor Model that underlies the NEO–PI–R emerged out of an impressive body of literature that has demonstrated that variables used to characterize normal personality consistently reduced to five domains when factor analyzed (Goldberg, 1993). Each of the five domain scales is divisible into six facet scales, representing major subdomains for each factor. In fact, the constructs underlying the NEO–PI–R facet scales are often little more than the test developers' hunches about which aspects of the domains are most important. Even so, it is the experience of many who use the instrument practically that the domain scales are too broad to be interpreted easily, and for descriptive purposes, the facet scales can be far more useful even if less reliable than the domain scores. The interpretation of these facet scales can in turn be complicated by the inclusion of disparate item contents. To cite just one example, the Feelings facet scale includes items having to do with the intensity of the respondent's emotional experience as well as items having to do with the willingness to be guided by emotions, which would seem experientially distinct phenomena.

Similarly, interpretive manuals for the MMPI describe a large body of evidence demonstrating the validity of the standard clinical scales, but then acknowledge it is often difficult to interpret the clinical scales because of their complexity, and recommend modifying the interpretation based on the review of far less validated subscales and even individual items (e.g., Greene, 2000). Similar advice may be found in texts discussing the clinical interpretation of complex neuropsychological and intelligence tests (e.g., Kaufman & Lichtenberg, 1999; Lezak, 1995). There are people who are sentimental but not warm, people who are sociable but not social. These elements of a person's functioning are perhaps at least as important to clinical description as global extravertedness or tolerance.

Issues of scaling. The extent to which multi-item scales provide an accurate placement for the respondent is another issue of little importance in the context of prediction, in which the primary concern is the demonstration that scale scores covary monotonically with relevant criteria. It is a matter of much greater importance when the scale is intended to represent the construct accurately. Standardization is the strategy most often used to address the problem of arbitrary scaling under the assumption that distance from the normative mean in standard deviation units can be used to estimate status. Apart

from the tremendous resources required to gather a normative sample—resources that could be devoted to more interesting research questions—standardization does not resolve the issue. For example, on many inventories such as the MMPI and NEO–PI, males and females produce different mean scores on many scales. Norming the test separately for males and females implies these differences are scale artifacts rather than evidence of gender differences in the constructs being measured. Ignoring other possible correlates of test performance such as age, socioeconomic status, and ethnicity implies that group differences on these variables are meaningful in terms of placement on the construct. Unfortunately, the lack of correspondence between scale and construct makes it difficult to be sure which interpretation is correct. It is possible to test these assumptions, for example, through the use of moderated regression analysis (Lautenschlager & Mendoza, 1986), but such analyses are rarely conducted, they add substantially to the complexity of standardization, and their accuracy depends on the construct validity of the criterion variables used (Bernstein, Teng, & Grannemann, 1987). The practice of standardizing to mitigate problems of placement is in turn compromised by the lack of construct validity in both the scale and relevant criteria.

Cohen, Cohen, Aiken, and West (1999) suggested the ambiguous meaning of scores on a multi-item scale is a major impediment to the development of precision in psychology as a science. Without intrinsically meaningful scaling, hypotheses about the expected direction of effects may be easy to justify; hypotheses about the expected size of those effects are not. Here too, standardization has been proposed as a means of addressing the problem. Meta-analyses of standardized effect size measures such as *r* or *d* are increasingly used to provide information about the size of effects. However, consider the following quote from Tukey (1969):

> I find the correlation coefficient a dangerous symptom. … What usually remains constant … is one of the regression coefficients. If we wish to seek for constancies, then, regression coefficients are much more likely to serve us than correlation coefficients.
>
> Why then are correlation coefficients so attractive? Only bad reasons seem to come to mind. Worst of all, probably, is the absence of any need to think about units for either variable. Given two perfectly meaningless variables, one is reminded of their meaninglessness when a regression coefficient is given since one wonders how to interpret its value. A correlation coefficient is less likely to bring up the unpleasant truth—we *think* we know what $r = -.7$ means. *Do we?* How often? Sweeping things under the rug is the enemy of good data analysis. … Being so disinterested in our variables that we do not care about their units can hardly be desirable. (p. 89)

Finding that psychological and educational interventions on average are a little less than ½ *SD* more effective than placebo (Lipsey & Wilson, 1993) seems to be better than simply stating an effect exists but in what way? It is unclear how this represents a meaningful improvement in understanding the impact of treatment. It is not even clear whether a difference of ½ *SD* has a consistent meaning across studies, the range of initial states, or research settings. In the absence of any evidence that this is so, researchers have come to rely heavily on benchmarks Cohen (1988) proposed for small, medium, and large effects. However, Cohen was clear that these were never intended as universal standards for the comparison of effect sizes and hoped that over time, the meaning of an effect size in a given context would become clearer. Seventeen years later there is no reason to believe Cohen's hope will ever be realized for standardized effect size statistics or whether it ever can be. Without an objectively meaningful benchmark, in many cases, a statement about the mean size of an effect should do little more than reinforce the conclusion that an effect exists.

SOME PRESCRIPTIVE SPECULATIONS

Clearly, both conceptually complex constructs and scales have an important role to play in psychosocial science. Conceptually complex constructs offer a useful means of summarizing information for purposes of communication and social policy. Conceptually complex scales are useful for predictive purposes. In cases in which conceptually complex constructs are socially useful, complex scales can also provide an index variable of the construct.

The basis of science is accurate measurement, though, and without accurate and precise measures, science can only advance so far. The focus on broadly defined constructs compromises the ability to develop precise models of psychosocial constructs. The focus on scales combining multiple contents compromises the ability to make precise statements about the meaning of a measurement.

The first step in improving the quality of measurement in psychology would require recognizing that the tendency to use the same measures for predictive and representational purposes is not optimal. The basic building blocks of measurement should demonstrate a high degree of representational accuracy. The discussion to this point would suggest that representation is maximized by combining highly redundant items representing narrowly defined constructs. More molar measures that are useful as predictors and index variables can then be built from these more molecular representational measures. However, these measures need not be assumed representative of any coherent construct. Prediction occurs in the context of operationism in that the goal of successful measurement is solely the maximization of an observable relationship, not the potential for intuitive understanding of the scale.

Some examples from the domain of performance testing can be used to demonstrate the issues involved in improving measurement practice. The structure of the SAT Rea-

soning test demonstrates the compromises that result from failing to distinguish between the predictive and representational goals of measurement. If the primary purpose of the instrument is to provide a predictor of college performance based on intellectual skills, it is likely, given Equation 1, that predictive validity would peak after administering three to four items representing each of four to five levels of difficulty. Further increments in predictive validity, if possible, would probably require broadening the range of skills sampled rather than adding more of the same types of items. If the primary purpose instead is to provide an accurate representation of mathematical or reasoning ability, these constructs are too broad to be represented accurately by any single score. Why then has the SAT evolved in the manner it has? The most likely reason would seem to be that restricting the test to these domains makes it easier to understand what each section of the examination is intended to represent, even if it does not represent the construct well. What results is a measure that is neither useful as an indicator of skills in these areas nor an efficient predictor of collegiate academic performance.

An example of the manner in which representational and predictive goals can be integrated more effectively is offered by the Woodcock–Johnson Tests of Cognitive Abilities (Woodcock, McGrew, & Mather, 2001). Each scale of the instrument is geared to the measurement of a single cognitive ability derived from a popular model of intellectual abilities (Carroll, 1993). The items on each scale are conceptually redundant, although they differ in item difficulty. Accordingly, the results for each scale can be meaningfully interpreted in terms of a single ability. For predictive purposes, those subscales that are relevant to each criterion can be aggregated. Research has suggested that different criteria call for different aggregates (e.g., Floyd, Evans, & McGrew, 2003), and there is no expectation that those aggregates reflect a single underlying construct.

Among multi-item rating scales, some scales such as the State–Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970) incorporate a fairly high proportion of conceptually redundant items. Even though scores on this measure often correlate quite highly with measures of other forms of psychopathology (e.g., McGrath & Ratliff, 1993), this is likely to suggest something about anxiety (although only if the other measure is also highly redundant).

One of the intriguing features of ratings is that even a single item can offer a very effective representation of the respondent's phenomenological experience. Consider the following item as an indicator of the lack of a sense of pleasure or enjoyment, a diagnostic criterion for depression:

In the past month, I believe I

A. have enjoyed my activities at least as much as the typical person.
B. have enjoyed activities a little less than the typical person would.
C. have experienced much less pleasure than the typical person does.
D. have had no or almost no pleasure or enjoyment from my activities.

This item is intended for illustrative purposes only and could probably be improved. It might be possible to identify response alternatives more consistent with theoretically meaningful steps in the experience of pleasure or a time frame that is more relevant to its natural history. The most serious problem is the degree to which the item calls for evaluation by the respondent of the meaning of concepts such as the typical experience of pleasure, an issue that is inherent to any linguistic interaction. Feedback from individuals knowledgeable about the experience, including both professionals and members of target populations, would be useful as a method of refining the item.

Even so, the response to this item by itself provides a correct and directly meaningful statement about how the respondent is willing to characterize his or her experience of pleasure relative to other individuals in the stated time frame. The outcome on this single item is an inherently interpretable datum that does not need combination with other items to reveal something that is potentially important about the individual's current status.

Generalizing from these examples, rating scales can provide the basis for construct valid measures if certain conditions are met:

1. The target construct is narrowly and precisely defined. The more elemental the construct, the more accurately it can be represented by a single graded scale.
2. The target construct represents an aspect of the respondent's experience of self or the environment so that a rating is the optimal mode of measurement. For example, although self-reported experience can be predictive of other types of constructs, there are many constructs for which self-report is incapable of maximizing the level of representational validity. Many of the key elements of the complex constructs narcissism, job performance, and wisdom have to do with perceptions by others or performance rather than experience. At times it is interesting to see whether self-perceptions match perception of the individual by clinicians or significant others, but these do not measure the same thing.
3. Items are developed so they closely match the theoretical conceptualization of the construct. This places an obligation on the test developer to provide such a conceptualization including information about the time frame or circumstances under which change in the construct may be expected, a list of reasonable gradations or discrete classes to serve as anchors for response alternatives, and the optimal mode of mea-

surement. This level of detail is probably only possible for specific constructs. The selection of response alternatives that reflect the theoretical structure of the construct argues against the common practice of using a single set of response alternatives for all items in a scale.

4. Controls are put in place to minimize the potential for error in measurement. This may include, for example, the evaluation of response styles, although there is increasing evidence that the importance of response styles has been overstated, at least in some settings (e.g., McGrath, Rashid, Hayman, & Pogge, 2002; Ones & Viswesvaran, 1998; Piedmont, McCrae, Riemann, & Angleitner, 2000).
5. Concerns about reliability can be addressed by administering two to three redundant items. This might involve administration of semantically related items, which would allow manipulation of item difficulty by shifting location along a dimension of intensity (e.g., "I feel sad" vs. "I feel morose" or shifting of response alternatives). Alternatively, averaging multiple administrations of a single item would maintain the meaningful scaling of the item. The administration of redundant items also allows for scale reversal, which can be used to detect yea-saying and nay-saying response styles (Barnette, 2000).

CONCLUSIONS

Several examples can be provided of the benefits that could result from the development of measures emphasizing representational accuracy. Despite over 50 years of research on response styles, the identification of overreporting and underreporting remains an unresolved dilemma. The model of testing I described in this article offers a fresh perspective on the issue. First, if modes of functioning involve different but interacting processes, it does not make sense to assume a true level of depression or job skill that some people misrepresent; instead, there are different opinions concerning the target individual's status on these variables. Second, the focus on individual items that are strongly representative of specific constructs will allow a much more fine-grained understanding of such differences in opinion. Interviews with respondents and independent raters could potentially identify circumstances under which self-ratings are more useful than ratings by others, circumstances under which ratings by others are more useful than self-ratings, and methods of identifying each. Similarly, interviewing respondents about changes in self-reports on specific items over time could provide useful information about the natural evolution of psychosocial states.

The creation of meaningfully scaled ratings also offers a statistical alternative to the reliance on standardized measures of effect size based on the characterization of groups of individuals at each interval on the scale. Consider a statement such as "36% of individuals who reported they were experiencing much less pleasure than the typical person during the month prior to initiating treatment reported a return to normal in the month following completion of cognitive-behavioral treatment compared with 18% of individuals exposed to placebo." It is admittedly more complex than a statement based on the size of *d,* but it has several advantages. It provides much more specific information about the effect. It is also inherently meaningful. One potential benefit of these attributes is that the reader is much more likely to be able to generate reasonable hypotheses about why this effect occurred. Having the ability to make intelligent guesses about why effects are of a certain size can only contribute to the maturation of psychology as a science. Furthermore, such statements avoid the representational problems created by the assumption of a quantitative structure to the experience of pleasure (Michell, 2000, 2001).

Finally, the proposed measurement model offers the potential for more efficient predictive methods. The administration of a standardized 20-item, self-report scale to predict some variable that can be predicted with equal accuracy by 4 items is wasteful. The benefit can be obscured by standard statistical practice, which treats the score on the full scale as a single predictor, resulting in the following regression equation:

$$Y = b_0 + b_1X_1.$$

When compared to the equation involving four individual items,

$$Y = b_0 + b_1X_1 + b_0 + b_2X_2 + b_0 + b_3X_3 + b_0 + b_4X_4,$$

the former can seem more parsimonious. The truth, though, is that the former equation is more accurately compared to the latter by expanding it as

$$\begin{aligned} Y &= b_0 + b_1(X_1 + X_2 + \ldots X_{20}) \\ &= b_0 + b_1X_1 + b_1X_2 + \ldots b_1X_{20}. \end{aligned}$$

The four-predictor equation does require estimation of five predictors rather than two, but given a large body of evidence indicating equal weighting of the predictors does not impact on predictive accuracy substantially (e.g., Dawes, 1979), the additional estimates may be considered optimal but unnecessary. If self-report measurement is restricted to three to four primary experiential factors, combined with one to two predictors from other relevant modes, there is the potential for a meaningful improvement in predictive accuracy even though the total number of predictors gathered has actually been reduced.

Despite dramatic advances in the understanding of psychosocial phenomena in recent years, the continued emphasis on complex constructs and scales remains a potential limiting factor to precision in psychosocial model building. I offer this article as the first step in the process of pursuing an alternate strategy to scale development built around the impor-

tance of maximizing representational validity through an emphasis on simpler indicators measuring simpler constructs.

ACKNOWLEDGMENTS

Portions of this article were presented at midwinter meetings of the Society for Personality Assessment in New Orleans, LA in 1999 and Albuquerque, NM in 2000. I am grateful to Deborah Bernstein, Donald Bernstein, and Gregory Meyer for their comments on earlier drafts of this article.

REFERENCES

Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2002). Ten-year comparisons of problems and competencies for national samples of youth: Self, parent and teacher reports. *Journal of Emotional and Behavioral Disorders, 10,* 194–203.

Acton, G. S. (1998). Classification of psychopathology: The nature of language. *The Journal of Mind and Behavior, 19,* 243–256.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Ashton, M. C. (1998). Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior, 19,* 289–303.

Ashton, M. C., Jackson, D. N., Paunonen, S. V., Helmes, E., & Rothstein, M.G. (1995). The criterion validity of broad factor scales versus specific facet scales. *Journal of Research in Personality, 29,* 432–442.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77,* 836–874.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60,* 361–370.

Bernstein, I. H., Teng, G., & Grannemann, B. D. (1987). Invariance in the MMPI's component structure. *Journal of Personality Assessment, 51,* 522–531.

Beutler, L. E., & Malik, M. L. (Eds.). (2002). *Rethinking the DSM: A psychological perspective.* Washington, DC: American Psychological Association.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117,* 187–215.

Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology, 80,* 168–177.

Broughton, R. (1990). The prototype concept in personality assessment. *Canadian Psychology, 31,* 26–37.

Burisch, M. (1984a). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39,* 214–227.

Burisch, M. (1984b). You don't always get what you pay for: Measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality, 18,* 81–98.

Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality, 11,* 303–315.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI–2: Minnesota Multiphasic Personality Inventory–2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Chisholm, S. M., Crowther, J. H., & Ben-Porath, Y. S. (1997). Selected MMPI–2 scales' ability to predict premature termination and outcome from psychotherapy. *Journal of Personality Assessment, 69,* 127–144.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstances for POMP. *Multivariate Behavioral Research, 34,* 315–346.

Collier, A. (1994). *Critical realism: An introduction to Roy Bhaskar's philosophy.* London: Verso.

Cone, J. D. (1995). Assessment practice standards. In S. C. Hayes, V. M. Follette, R. M. Dawes, & K. Grady (Eds.), *Scientific standards for psychological practice: Issues and recommendations* (pp. 201–224). Reno, NV: Context.

Cooper, R. (2004). What is wrong with the *DSM*? *History of Psychiatry, 15,* 5–25.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO–R) and NEO Five-Factor Inventory (NEO–FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Crowley, S. L., & Merrell, K. W. (2000). Convergent and discriminant validity of the Internalizing Symptoms Scale for Children. *Journal of Psychoeducational Assessment, 18,* 4–16.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.

Edelmann, R. J., & Baker, S. R. (2002). Self-reported and actual physiological responses in social phobia. *British Journal of Clinical Psychology, 41,* 1–14.

Ehrenworth, N. V., & Archer, R. P. (1985). A comparison of clinical accuracy ratings of interpretive approaches for adolescent MMPI responses. *Journal of Personality Assessment, 49,* 413–421.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51,* 360–392.

Eysenck, H. J. (1947). *Dimensions of personality.* London: Routledge & Kegan Paul.

Floyd, R. G., Evans, J. J., & McGrew, K. S. (2003). Relations between measures of Cattell–Horn–Carroll (CHC) cognitive abilities and mathematics achievement across the school-age years. *Psychology in the Schools, 40,* 155–171.

Fried-Buchalter, S. (1992). Fear of success, fear of failure, and the imposter phenomenon: A factor analytic approach to convergent and discriminant validity. *Journal of Personality Assessment, 58,* 368–379.

Friedman, M. (1991). The re-evaluation of logical positivism. *Journal of Philosophy, 88,* 505–519.

Gergen, K. J. (1985). The social constructionist movement in modern psychology. *American Psychologist, 40,* 255–265.

Ghiselli, E. E. (1964). *Theory of psychological measurement.* New York: McGraw-Hill.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34.

Gotlib, I. H., & Cane, D. B. (1989). Self-report assessment of depression and anxiety. In P. C. Kendall & D. Watson (Eds.), *Anxiety and depression: Distinctive and overlapping features* (pp. 131–169). San Diego, CA: Academic.

Greene, R. L. (2000). *The MMPI–2: An interpretive manual* (2nd ed.). Boston: Allyn & Bacon.

Hamilton, M. (1967). Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology, 6,* 278–96.

Hampson, S. E., John, O. P., & Goldberg, L. R. (1986). Category breadth and hierarchical structure in personality: Studies of asymmetries in judgments of trait implications. *Journal of Personality and Social Psychology, 51,* 37–54.

Hogan, J., & Roberts, B. W. (1996). Issues and non-issues in the fidelity-bandwidth trade-off. *Journal of Organizational Behavior, 17,* 627–637.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology, 83,* 693–710.

Kagan, J. (1988). The meanings of personality predicates. *American Psychologist, 43,* 614–620.

Kaufman, A. S., & Lichtenberg, E. O. (1999). *Essentials of WAIS–III assessment.* New York: Wiley.

Kirsch, I., Moore, T. J., Scoboria, A., & Nicholls, S. S. (2002, July 15). The emperor's new drugs: An analysis of antidepressant medication data submitted to the U.S. Food and Drug Administration. *Prevention & Treatment, 5,* Article 23. Retrieved September 25, 2002, from http://journals.apa.org/prevention/volume5/pre0050023a.html

Lautenschlager, G. J., & Mendoza, J. C. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement, 10,* 133–139.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist, 48,* 1181–1209.

Lorr, M. (1986). *Interpersonal Style Inventory (ISI) manual.* Los Angeles: Western Psychological Services.

Mahrer, A. R. (1999). Embarrassing problems for the field of psychotherapy. *Journal of Clinical Psychology, 55,* 1147–1156.

Manicas, P. T., & Secord, P. F. (1983). Implications for psychology of the new philosophy of science. *American Psychologist, 33,* 399–413.

McGrath, R. E., Rashid, T., Hayman, J., & Pogge, D. L. (2002). A comparison of MMPI–2 high-point coding strategies. *Journal of Personality Assessment, 79,* 243–256.

McGrath, R. E., & Ratliff, K. G. (1993). The use of self-report measures to corroborate theories of depression: The specificity problem. *Journal of Personality Assessment, 61,* 156–168.

Meehl, P. E. (1945). The dynamics of "structured" personality tests. *Journal of Clinical Psychology, 1,* 296–303.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244.

Meehl, P. E. (1995). Utiles, hedons, and the mind-body problem, or, who's afraid of Vilfredo? In P. E. Shrout & S. Fiske (Eds.), *Personality research, methods, and theory: A Festschrift for Donald Fiske* (pp. 45–66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mershon, B., & Gorsuch, R. L. (1988). Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria? *Journal of Personality and Social Psychology, 55,* 675–680.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance 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.

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist, 36,* 211–217.

Millon, T. (1977). *MCMI–II manual.* Minneapolis, MN: National Computer Systems.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

Neuberg, S. L., West, S. G., Judice, T. N., & Thompson, M. M. (1997). On dimensionality, discriminant validity, and the role of psychometric analyses in personality theory and measurement: Reply to Kruglanski et al.'s (1997) defense of the Need for Closure Scale. *Journal of Personality and Social Psychology, 73,* 1017–1029.

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

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11,* 245–269.

Paunonen, S. V. (1998). Hierarchical organization of personality and prediction of behavior. *Journal of Personality and Social Psychology, 74,* 538–556.

Paunonen, S. V., & Jackson, D. N. (1985). The validity of formal and informal personality assessments. *Journal of Research in Personality, 19,* 331–342.

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78,* 582–593.

Prescott, C. A., & Gottesman, I. I. (1993). Genetically mediated vulnerability to schizophrenia. *Psychiatric Clinics of North America, 16,* 245–267.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology, 4,* 328–350.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.

Sayer, N. A., Sackeim, H. A., Moeller, J. R., Prudic, J., Devanand, D. P., Coleman, E. A., et al. (1993). The relations between observer-rating and self-report of depressive symptomatology. *Psychological Assessment, 5,* 350–360.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350–353.

Schumaker, R. E., & Lomax, R. B. (1996). *A beginner's guide to structural equation modeling.* Hillsdale NJ: Lawrence Erlbaum Associates, Inc.

Simpson, R. H. (1944). The specific meanings of certain terms indicating differing degrees of frequency. *Quarterly Journal of Speech, 30,* 328–330.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *The State–Trait Anxiety Inventory (STAI): Test manual.* Palo Alto, CA: Consulting Psychologists Press.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco: Jossey-Bass.

Taylor, J. B., Ptacek, M., Carithers, M., Griffin, C., & Coyne, L. (1972). Rating scales as measure of clinical judgment. 3: Judgments of the self on personality inventory scales and direct ratings. *Educational and Psychological Measurement, 32,* 543–557.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 33,* 83–91.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Reading MA: Addison-Wesley.

Woodcock, R., McGrew, K., & Mather, N. (2001). *Woodcock–Johnson Battery–Third Edition (WJ–III).* Itasca, IL: Riverside.

Robert McGrath
School of Psychology
T–WH1–01
Fairleigh Dickinson University
Teaneck, NJ 07666
Email: mcgrath@fdu.edu

Received May 3, 2004
Revised October 25, 2004