Psychological Assessment
1998, Vol. 10, No. 4, 316-330

Foundations of Psychological Assessment: Implications for Cognitive Assessment in Clinical Science

Richard M. McFall and James T. Townsend
Indiana University Bloomington

We examine psychological assessment within the broader framework of psychology's efforts to build and test useful scientific theories. In the first section, we consider in detail a number of fundamental epistemological, conceptual, and methodological issues that tend either to inhibit or to foster theoretical progress in psychology. In light of these issues, we then recommend that psychology adopt an information-based, quantitative approach to theory building and testing. This approach should help us model the dynamic, stochastic processes underlying human behavior. In the second section, we explore the implications of the issues and strategies that we outlined in the first section for the future of clinical assessment, with a particular focus on the clinical assessment of cognitive processes. We conclude by advocating a conceptual and methodological integration of clinical and cognitive neuroscience in psychology.

This article critically examines conceptual and methodological issues in psychological assessment, and offers prescriptions for future improvements. It is divided into two sections. The first and the longest section reviews the foundations of assessment, as we see them. Because all evaluation and advice is based on preconceptions about how best to differentiate good from bad, throughout this section we lay bare the perspective and assumptions that guided our analysis and advice. The first section pertains to assessment in general; in the second section, this material is applied to an evaluation of assessment in clinical psychology, with a particular focus on cognitive assessment.

We do not equate assessment with psychological tests and measures, in the narrow sense that has become commonplace in clinical psychology. We construe assessment in a broader and more fundamental way—as an integral part of the scientific enterprise, whether focused on basic or applied questions. We present this view of assessment in the first section. Although some of this material is basic, it warrants review. Many current shortcomings in clinical assessment reflect investigators' inattention to such fundamentals. Some other material may seem more difficult and remote, but its significance should become apparent as things unfold. We hope that readers will accept, at least provisionally, the structure we build in the first section.

In the second section, we use the structure and tools laid out in the first section to examine critically the current status of assessment in clinical psychology, with a particular focus on cognitive assessment. Our critique is succinct and selective, not exhaustive. It focuses on general conceptual and methodological issues, rather than on specific theoretical questions, tests and measures, or experimental results. Our criticisms are generic. The goal is to illustrate common strengths and weaknesses in clinical assessment, not to criticize individuals; examples are cited only to help make general points. To make our criticisms clear, we contrast clinical psychologists' concepts and methods with those of contemporary cognitive scientists. These comparisons do not imply that the latter group is without serious unresolved problems of its own; the comparisons simply bring into sharper focus some of the key issues that clinical scientists need to consider if they hope to advance cognitive assessment.

The overall aim of this article is to promote improvements in psychological assessment, especially in the clinical assessment of cognition. Thus, where possible, we have illustrated our points by citing examples at the cutting edge of psychological research in cognitive and clinical science, the areas we know best. Clinical scientists who otherwise might wish to sharpen their approach to assessment may not see clearly how to translate our abstractions into concrete actions within their own research area. By pointing to a handful of exemplars, we hope to prime the flow of creative ideas essential to new problem-specific translations. This article concludes with several caveats aimed at dispelling any illusions that we think either (a) that cognitive assessment is simple, or (b) that we have provided all of the answers or solved all of the problems.

Foundations of Psychological Assessment

All assessments are not created equal. They differ in at least three critical ways. First, they differ in terms of the theoretical questions they are designed to illuminate. Second, they differ in their logical structure. Third, they differ in the quality of the information they yield and, hence, the weight of the inferences that they can support. In this first section, we explore these three differences and consider their implications for psychological assessment in general.

Theoretical Questions

Theories as models. The purpose of all scientific assessment is to shed light (i.e., to provide information; to reduce uncer-

Richard M. McFall and James T. Townsend, Department of Psychology, Indiana University Bloomington.

Correspondence concerning this article should be addressed to Richard M. McFall, Department of Psychology, Indiana University, Bloomington, Indiana 47405. Electronic mail may be sent to mcfall@indiana.edu.

tainty [Mischel, 1968; Shannon & Weaver, 1949]) on specific questions, from basic to applied. Thus, there is an essential link between theory and assessment: Assessments must be tailored to fit the specific theoretical questions they are designed to answer (McFall, 1993; Townsend, 1975, 1994).[1]

Science progresses best, in our view, by building and testing models of nature (Polya, 1957; Popper, 1962). Thus, scientific theories are representational constructions. These constructions focus selectively on a few abstracted dimensions, or patterns, of similarity and difference in nature, thereby ignoring the multitude of other potential abstractions (Kelly, 1955; Polya, 1957). When fully developed, these models are multilayered, as depicted in Figure 1.

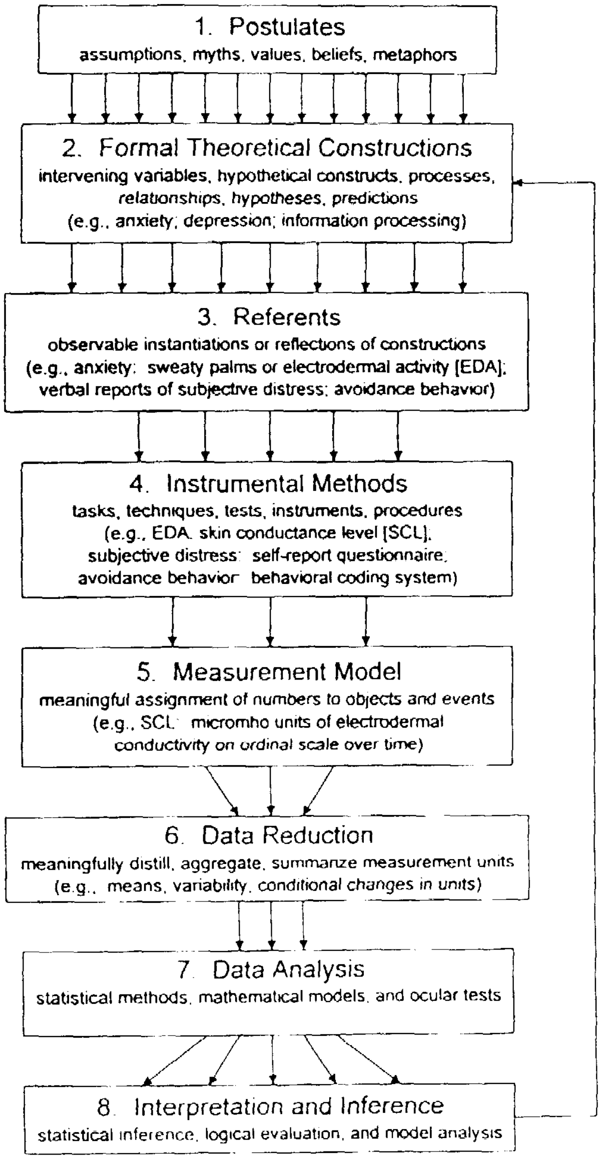
1. Postulates
assumptions, myths, values, beliefs, metaphors
2. Formal Theoretical Constructions
intervening variables, hypothetical constructs, processes, relationships, hypotheses, predictions
(e.g., anxiety; depression; information processing)
3. Referents
observable instantiations or reflections of constructions
(e.g., anxiety: sweaty palms or electrodermal activity [EDA]; verbal reports of subjective distress; avoidance behavior)
4. Instrumental Methods
tasks, techniques, tests, instruments, procedures
(e.g., EDA: skin conductance level [SCL]; subjective distress: self-report questionnaire; avoidance behavior: behavioral coding system)
5. Measurement Model
meaningful assignment of numbers to objects and events
(e.g., SCL: micromho units of electrodermal conductivity on ordinal scale over time)
6. Data Reduction
meaningfully distill, aggregate, summarize measurement units
(e.g., means, variability, conditional changes in units)
7. Data Analysis
statistical methods, mathematical models, and ocular tests
8. Interpretation and Inference
statistical inference, logical evaluation, and model analysis

Figure 1. Eight layers of a scientific model building and testing process.

The outermost layer consists of *postulates* (i.e., philosophical assumptions, myths, values, beliefs, metaphors) that serve as the model's exoskeletal structure (Polya, 1957; Smith, 1984). This assumptive layer is impervious to logical or empirical proof, hence beyond immediate scientific test. The postulates simply are treated as "givens." Once stipulated, however, they constrain all that follows, sometimes in subtle ways (Lakoff & Johnson, 1980).

The next layer consists of *formal theoretical constructions* (i.e., intervening variables; hypothetical constructs and their nomological networks; theoretical mechanisms and processes; and explicit statements of relationships, hypotheses, and predictions [Cronbach & Meehl, 1955; MacCorquodale & Meehl, 1948]). This theoretical layer must be congruent with the model's postulates. However, so many plausible theories satisfy this criterion that scientists must look to deeper layers of their models to resolve questions of truth, or to choose among the multitude of competing theories. Anxiety is an example of a hypothetical construct at this theoretical level. (Note that at the postulate layer, anxiety is represented metaphorically by such expressions as "high strung" [a violin string?], "nervous" [too many nerves firing too rapidly?], and "angst" [from a German word for "choking"]. It simply is assumed, without proof or justification, that feeling anxious is bad, something to be minimized or avoided.)

The next layer, the *referent* level, consists of observable events that give concrete meaning and definition to the constructs at the formal theoretical layer. For instance, the hypothetical construct of anxiety, at the theoretical level, is linked to several specific observable events, or instantiations, at the referent level, such as sweaty palms, trembling hands, avoidance, and self-reported anxious feelings (Kozak & Miller, 1982; Lang, 1968).

To evaluate empirically the correspondence between theoretical expectations and real-world experiences, systematic samples of the observable referents must be collected by some method. Thus, it is at this next layer, the *instrumental methods* level, that scientists focus on such data-collection matters as tasks, techniques, tests, instruments, and procedures. It is at this level that we are most concerned with issues of representativeness, standardization, reliability, and contamination. A common method of sampling anxiety, for example, is to measure electrodermal activity (EDA), which provides data on the observable referent of sweaty palms through psychophysiological recordings of skin conductance levels over time and across conditions. But we could choose a different method to assess sweaty palms (blotters?) or could sample a different referent altogether. For instance, we might use a paper-and-pencil, self-report method to sample subjective experiences of anxiety, or we might use an observer coding method to sample approach-avoidance behavior. Each of these methods would provide a different sample of the theoretical construct, anxiety. If the construct is a good one, these different sampling methods should yield convergent evidence, and their results also should diverge from those of methods that sample referents for other constructs (Campbell & Fiske, 1959).

[1] Also see *Philosophy and the Science of Behavior* by Merle B. Turner (Turner, 1967), an underappreciated, thoughtful, extensive, and durable treatment of philosophy of science in psychology.

Although events at the referent level may be more objective and observable than the theoretical constructs they instantiate, they must be linked formally, in turn, to units at the next layer, the *measurement model* layer, before they can be useful. It is at this level that issues of measurement scale and instrument calibration become critical. For a specific theory to be tested, its constructs, referents, and sampling methods must be translated into a quantitative measurement model that represents meaningfully the relationships between constructs and their sampled referents.[2] "Measurement is (or should be) a process of assigning numbers to objects in such a way that interesting qualitative empirical relations among the objects are reflected in the numbers themselves as well as in important properties of the number system" (Townsend & Ashby, 1984, p. 394).[3] Axiomatic measurement theory offers a number of proven theorems regarding the fundamental requirements for scientific measurement models. For instance, it clearly specifies the unique properties of different measurement scales (nominal, ordinal, interval, ratio), and explicates the statistical, inferential, and theoretical implications of each. We will have more to say about measurement theory later; however, because space limitations prevent us from covering the subject in detail, unfamiliar readers are urged to pursue it on their own (Krantz, Luce, Suppes, & Tversky, 1971; Roberts, 1979; Townsend, 1990b; Townsend & Ashby, 1983, 1984).

Continuing with our example of anxiety, palmar sweat (a referent for anxiety) typically is assessed through psychophysiological methods of measuring EDA. These methods, in turn, yield quantitative indexes of EDA, such as skin conductance level (SCL). Thus, at the measurement level, palmar sweat gland activity is translated into micromho units of electrodermal conductivity on an ordinal scale over time. SCL is not a direct measure of sweat gland activity per se, but an indirect measure, a correlate, or a sign of such activity. These signs of palmar sweat, then, are the actual units that we use to test our abstract theories of anxiety, arousal, or emotion. When investigators choose instead to assess anxiety by a self-report method, their measurement model typically transforms subjective experiences of anxiety into ordinal scale numbers reflecting participants' pencil marks on a paper questionnaire, again, a sign. These uses of "signs," as opposed to "direct samples," in measurement models are neither bad nor good intrinsically. They offer potential advantages, such as increased standardization, objectivity, replicability, and quantitative rigor. But they also carry with them the potential danger of seducing investigators into reifying their measurement units, as though these were the real phenomenon of interest, rather than indirect reflections of it.

The use of signs raises a deeper theoretical question as well: What are the processes or mechanisms by which these signs are related to the phenomena of interest? As long as the relationship is purely correlational, it has little explanatory value. Although a sign may have predictive value, this value is inherently tenuous. As long as we do not know why things are related, we cannot assume that the relationship will last, nor can we predict the circumstances under which it might evaporate. Nevertheless, if these relationships seem stable, we often are tempted not only to speculate about their causes, but also to believe in our own speculations. We need to guard against such self-deception. We should recognize that signs are place-holders, used in the absence of sound theoretical explanations. If we understood the underlying causes, we would refer to these processes, rather than to the signs.

At the subsequent level, the *data reduction* level, the measurement units generated by our instrumental methods must be distilled, aggregated, and summarized. For example, SCL is represented by a continuous electrical output signal of micromho units of conductivity, displayed graphically as squiggly lines on a paper chart or an oscilloscope. These raw data can be distilled into any number of different indices (e.g., signal amplitude values, such as overall mean or period-by-period mean; signal variability values, such as range, standard deviation, absolute or percentage change from baseline in amplitude, frequency, or slope; or derivative conditional values, such as stimulus-dependent changes in key indices). Scientists must decide which indices, or data reduction strategies, best capture (or reflect) the intended meaning of the measurement units chosen to represent the referents for the theoretical constructs. Once again, fundamental measurement considerations are critical. The choice of measurement units and scales necessarily constrains the range of reasonable data-reduction options. Data-reduction strategies should be congruent with the concepts, decisions, and structures at all superordinate levels of the model. And the choice of measurement models and data-reduction strategies, in turn, will limit the degrees of freedom available at all lower layers of the modeling process.

The next layer of scientific modeling consists of *data analysis*. Psychologists traditionally have relied heavily on a hybrid mixture of Fisherian and Neyman–Pearsonian statistical methods for analyzing their data; however, this convention is only one of many potential approaches. Indeed, psychologists' commitment to these data-analytic methods is not shared by many scientists in other so-called hard sciences (Gigerenzer et al., 1989). Regardless of one's choice of data-analytic methods, however, it is important to realize that no method is atheoretical, assumption free, or without limitations. Indeed, all data analysis involves a quantitative modeling process that is inherently theoretical (i.e., specific to the question being addressed), conjectural (i.e., "suppose we look at it this way"), and conditional (e.g., dependent on one's method and measurement model; see Gigerenzer & Murray, 1987; Meehl, 1971). Currently, for example, there is considerable ferment and debate among psychologists regarding the value of the traditional null hypothesis significance testing approach to data analysis. As this debate reveals, there is no absolute, correct, "royal road to truth" in data analysis. Space limitations prevent a fuller discussion of this debate, but interested readers are referred to Cohen (1994), to Loftus (1996), and to a special section of *Psychological Science* (Harris, 1997) for an exposition and review.

[2] "Meaningfulness" is a technical, all-or-none term here. Scales either are appropriate to their theory and measurement process, or they are not; thus, "a statement can not be almost meaningful" (Townsend & Ashby, 1984, p. 395). However, some technically meaningful statements may convey more information, exert a greater influence, or be more useful than others.

[3] Qualitative relations are not necessarily incompatible with quantitative relations, for example, "greater than" is a qualitative statement about an empirical quantitative relationship.

The final layer of scientific modeling consists of the *interpretation and inference* process. Logical inferences and their theoretical implications are gleaned from the results of the data analyses. The quality and strength of the plausible inferences, at this point, are constrained logically by decisions made at all prior levels of the model (e.g., see Townsend, 1990b, for a theory of hierarchical inference related to measurement scale). These inferences and implications, in turn, feed back vertically through all superordinate layers of the model. When things go right, these inferences illuminate the questions that gave rise to the assessment process in the first place. But the inference process is not simple; there is no such thing as an "automatic inference machine" in science (see Meehl, 1971, for an explication of this point). In general, all logical inferences are constrained by our postulates, theories, and initial questions; by our choice of referents, sampling methods, and measurement models; and by our strategies for data reduction and analysis. Within these constraints, we seek new information that allows us to eliminate plausible rival hypotheses, and to replace our a priori theoretical models with a posteriori models that are more useful for describing, predicting, explaining, and controlling events.

Just as the tumblers in the combination lock on a safe must be aligned properly before one can gain access to the safe's contents, all eight layers of the scientific model depicted in Figure 1 must be aligned vertically and coherently before our assessments can yield meaningful answers to our theoretical questions (i.e., reduce our uncertainty about nature; Gigerenzer & Murray, 1987; Meehl, 1971). In science, of course, we do not know a priori the correct combination to nature's safe, or how best to align the layers of our model. We must search for the combination through a trial and error process of hypothesis testing. Indeed, this is the very business of science. Popper (1962) has identified this process of systematic "conjectures and refutations" as the distinguishing feature of a scientific epistemology.

Utility. The scientific merit of a constructed multilayered scientific model is a function of its utility. *Utility*, from this perspective, is a complex and relativistic yardstick, reflecting both the degree to which a model consistently (reliability) does what it was intended to do (validity), and the degree to which it does this better than competing models (incremental validity; strong inference; Feynman, 1985; Mischel, 1968; Platt, 1964).

All models have only limited utility because they represent only selected facets of nature; therefore, each model's scientific merit must be judged in relation to the limited domain staked out by the theory, defined by its claims, purposes, focus, and range of convenience. This implies that a particular model's merit cannot be determined if its domain has not been defined adequately beforehand.

This utilitarian view of theoretical models suggests a Bayesian epistemology. That is, a model has scientific value to the degree that it reduces uncertainty concerning focal questions within its limited domain (i.e., relative to a priori uncertainty, or to what was "known" before the model was proposed). Thus, optimally, a model's scientific utility should be evaluated quantitatively. This means that models that have been specified in terms that can be tested quantitatively are most likely to demonstrate scientific utility.

Quantitative models differ in their fundamental properties, including their scope and level of specificity; their underlying scale, their assumptions about distributions, variability, sampling, and determinacy; and their theoretical complexity and precision. These different properties have important and systematic implications for differences in the power of these models to describe, predict, explain, and control phenomena within their domain (Townsend, 1990b). The more rigorously specified the fundamental properties of a quantitative model, the more that model is at risk of being falsified; hence, to the extent that it survives attempts at falsification, the greater its utility, power, and stature as a scientific model (Popper, 1962).

If the most informative scientific assessments are designed to put our model's theoretical propositions at risk of falsification, then assessments not only should be capable of assigning meaningful values to the theory's unique constructs, referents, and measures, but they also should permit a quantitative evaluation of the degree of convergence and divergence, or fit, between the predicted and observed patterns of relationships. Assessments that satisfy these criteria allow us to evaluate a given theory's utility, or scientific merit. Essentially, scientific assessments provide quantitative representations of the observable referents for our theoretical constructions. These representations, in turn, allow us to quantify the degree of correspondence (at preselected points of contact) between empirical reality, on the one hand, and abstract theoretical expectations (hypotheses, predictions), on the other hand. With this evidence, we judge the utility of our scientific models.

Varieties of assessment questions. The specificity of assessment questions can vary considerably, from vague to precise. All assessment, however, even the most informal and exploratory, necessarily implies some kind of theoretical preconception. The events targeted for assessment, the referents sampled, the methods used, and the measurement units recorded must be selected, on some basis, from the universe of possibilities. These choices reflect the assessor's implicit and explicit preconceptions (hunches, expectations, theories), without which the assessment effort would be pointless.

The types of theoretical questions that assessments are designed to answer vary, of course. Most assessments in experimental psychological research are concerned with nomothetic questions; that is, they are designed to test specific theoretical predictions about human beings in general. Individual differences either are treated as error ("noise"), or are modeled as probability distributions. In contrast, assessments in clinical psychology are concerned more often with idiographic questions; that is, they are designed to guide clinicians in making decisions regarding diagnosis, etiology, prognosis, intervention, and outcome evaluation in individual cases or subgroups. Note, however, that meaningful answers to idiographic questions depend on the preexistence of answers to nomothetic questions. Background knowledge about underlying probability distributions, conditional probabilities, and functional relationships is essential to the assessment and interpretation of individual differences. This point has been made repeatedly and forcefully for nearly half a century (e.g., Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996; Meehl, 1954), yet it remains elusive to many clinicians (e.g., Matarazzo, 1990; D. R. Peterson, 1996; cf. McFall, 1996). All assessment and prediction is fundamentally nomothetic; without some knowledge about the expected probability distribution of values on a given measure (whether the measure be an informal interview or a formal test), an isolated value on that measure is uninterpretable.

The simplest assessment questions are *descriptive* (e.g., What is x? How many x are there? How often does x occur?). These typically can be addressed with simple methods that yield counting, naming, or yes–no data, and with simple measurement models that represent the data on nominal or ordinal scales. However, assessment questions that focus on *relationships* (e.g., What is the probability of x, given y? How is the value of x related to the value of y? How are changes in x related to changes in y?) are more likely to involve interval or ratio scales, which require more rigorous methods and more precise measurement models. The most complex theoretical questions, such as those seeking to map functional relationships with mathematical equations, almost certainly involve interval, ratio, or absolute scales, and require the most advanced methods and measurement models of all. Although it may be feasible to use less rigorous measurement models to address simpler questions, there is a hierarchy of inference: Stronger models always imply solutions to lower order questions, but weaker models cannot address higher order questions (Townsend, 1990b).

Logical Structure of Assessment

Many clinical psychologists mistakenly equate the instrumental methods used in assessment with the clinical assessment process itself, as though such methods had intrinsic meaning and independent value. Instrumental methods are an integral component of assessment, of course, but assessment is much more than methods alone. All eight layers of the modeling process are involved. Weakness at any layer or incongruence between layers undermines the assessment process. Meaningful assessment requires clear and logical relationships across the entire multilayered model.

With unfortunate consequences, psychologists too often have blurred the distinctions and logical relations among the different layers. One reason for this confusion may be that although the layers initially evolved hand in hand as the scientific model was developed, over time the original rationale for the distinctions and connections among the layers gradually tends to fade from memory. Sometimes this leads investigators to treat individual layers as though they were functionally autonomous (cf. Astin, 1961); at other times it leads them to treat conventional connections among layers as though they were sacrosanct. The resulting confusion leads psychologists into logical predicaments. We identify the following five common logical errors that deserve particular attention:

1. Acting as though an assessment method or data-analytic technique were a theory. This error is evident, for example, when investigators select assessment instruments on the basis of their names alone, with little thought to the theoretical and measurement models behind these instruments. Investigators who do this have jumped to several illogical conclusions: (a) that the instrument's name actually defines what it measures, (b) that the construct named by the instrument must be part of a coherent and valid theory, (c) that the theory behind the instrument is congruent with the investigator's theory, and (d) that any investigators who use the instrument automatically tap into the presumed validity of the instrument and the theory behind it.

In the case of data-analytic techniques, a parallel logical slippage is revealed, for example, in the explosive growth of "the normal curve" assumption, dating back to the end of the 19th century and beginning of the 20th century. Much of statistics was and is founded on the assumption of normality. Even today, it is practically obligatory to present data coded into standard deviates, regardless of the appropriateness of such transformations to the data at hand.

Another example, with a twist, is the popularity in the 1950s of using Shannon and Weaver's Information Theory (Shannon & Weaver, 1949) for analyzing the transmission of information in humans and animals. This theory and associated methodology delivered accuracy per-unit-time measures on a strong measurement scale (i.e., bits per unit of time). However, some investigators confused the measuring stick (i.e., the computation of "bits of information transmitted") with the superordinate theory. Thus, when George Miller (1956) demonstrated that the "bits of information" measure was not the critical limitation in short-term memory and absolute-judgment tasks, some investigators jumped to the conclusion that the superordinate information theory should be discarded. This was an error, of course, as information theory and its strong measurement scale remained potentially useful for a variety of purposes, Miller's demonstration notwithstanding (Kantowitz & Knight, 1976). It simply is a mistake to equate specific methods of measuring theoretical constructs with the overall theory itself.

Faddish, all-or-none thinking about methods and data-analytic techniques tends to foster an atmosphere of political correctness. Sometimes, investigators may feel pressured to use "in" techniques, even when these are inappropriate; at other times, investigators may discover that their work is being devalued because they used "out" techniques, even though these were used appropriately (see Meehl, 1971, for an extended discussion of this point in the context of a specific example). But assessment methods and data-analytic techniques are not theories, to be rejected or accepted on their own; rather, they exist to provide empirical evidence regarding specific theoretical models, and their value depends entirely on their ability to illuminate and test these models.

2. Acting as though a tool of measurement or analysis were assumption free. Too often, psychologists use assessment instruments and statistical procedures without considering their underlying assumptions or overhead costs (see Townsend, 1994). Whereas this error may seem almost the polar opposite of the previous error, it is no less problematic. Before using off-the-shelf assessment instruments and conventional statistical tests conveniently packaged in computer software, psychologists must stop to consider the appropriateness and implications of such choices. The less conventional the technique, the greater the care required. For example, techniques such as factor analysis, path analysis, structural equation modeling, and discriminant function analysis can be valuable tools when used cautiously and knowledgeably, yet all are constrained by underlying assumptions that can create serious problems if ignored. Factor analysis, in its elaborated form (beyond principal components), is susceptible to indeterminacy problems. Especially when used as part of a "fishing expedition" (in contrast to confirmatory factor analysis), its output simply cannot be interpreted with confidence (e.g., Floyd & Widaman, 1995). Path analysis and structural equation modeling rarely have been used with sufficient restrictions to evade identifiability problems. And the results of discriminant function analysis are uninterpretable without replication or, at the very least, without use of the often neglected jackknifing procedure (Lachenbruch & Mickey, 1968).

In essence, all statistical procedures are based on conceptual and measurement models, just as the measurements of force in physics or molecular weight in chemistry are parts of models. Because the behavioral sciences have no well-established models comparable to those in the hard sciences, we sometimes compensate by keeping our assumptions as innocuous as possible, hoping that simplicity will make our results more interpretable and the emergent principles more evident. This is a reasonable strategy, but not without risk. Occam's razor notwithstanding, there always is the possibility that seemingly innocuous assumptions may prove theoretically misleading, or even wrong. There is no escaping the constraints of one's underlying assumptions. Therefore, whether one's assumptions are simple or complex, they should be stated explicitly and their coherence across all layers of one's model should be examined carefully.

3. Disregarding questions of "strength of scale" in measurement and analysis. The generative idea of different scale types can be traced to the prescriptions of psychophysicist S. S. Stevens (e.g., 1946, 1951), who said that measurements vary according to what changes could be imposed on the measurement numbers themselves. The more constrained the changes of numbers representing the object measurements, the more numerical operations and statistical procedures that could be visited on these numbers lawfully. The central precept is that a statement (statistic, inequality, etc.) should be true whatever the permissible change to the representing number. Thus, any statement or formula involving numbers on an *absolute* scale (e.g., counting; probability; absolute [Kelvin] temperature) will be true invariably, because no alteration at all is permissible. A *ratio* scale (e.g., length, time, or mass) permits multiplication by a positive number as a change in unit (e.g., inches to centimeters), and any statement that uses ratios (only) will be true whatever the alteration in unit. For example, a mean still will carry the specific unit, but, say, a difference between means divided by the common standard deviation will cancel out the unit and make that statement invariantly true. An *interval* scale, in contrast, permits both a unit shift as well as a shift of origin to occur (e.g., Celsius vs. Fahrenheit temperature). Mean differences divided by the standard deviation would remain unchanged because in the numerical operations to derive them, any shift in origin also would be subtracted out; ergo, the probity in applying Z or t tests to measurements based on at least interval scale strength. However, a ratio of two means would be invariant for a ratio scale, but not an interval scale. Finally, the weakest standard scale is the *ordinal* scale (e.g., the hardness scale; perhaps most psychological scales). This scale permits any alteration that preserves order (i.e., any strictly increasing transformation), but most numerical statements are prohibited because their truth value would be affected by such alterations. However, numerical inequalities may be invariant and certain types of statistical comparisons based on inequalities will adhere to this demand (e.g., Townsend, 1990b).

Pioneers in the field of axiomatic measurement (e.g., Krantz et al., 1971; Roberts, 1979) have developed a rigorous theory of how numbers and operations in a numerical system (e.g., arithmetic) can reflect the properties of real-world objects and events (e.g., mass, intelligence, and speed of information processing). The concepts in these theories are critical to psychological assessment. In particular, these concepts make it clear that it is illegitimate to treat a measurement as though it were on a strong scale without first demonstrating that it actually is!

Unfortunately, psychologists too often ignore the scaling implications of their sampling methods, measurement models, data-reduction strategies, and data-analysis techniques. Consider, for example, psychologists' long-standing reliance on self-report questionnaires to assess individual differences in personality. These questionnaires usually sample participants' true–false responses to a large number of test items purportedly representing a common personality domain. The measurement model underlying such questionnaires typically makes three basic assumptions: (a) *item equivalence*: all items are assumed to be equally representative of the target personality characteristic; hence, they are treated as interchangeable and weighted equally; (b) *response additivity*: the more items a participant answers in the keyed direction, the more of the personality characteristic the participant has; thus, test scores are simply the sum of each participant's keyed responses; and (c) *ordinal scale*: there is an ordinal relationship between test scores and strength of the personality characteristic. Although this last assumption often is ignored, its implications are critical. Suppose that participants A, B, and C earned test scores of 5, 10, and 15, respectively. Assuming an ordinal scale, we can conclude only that C has more of the characteristic than B, and that B has more of it than A. We cannot say, for instance, that the difference between A and B is equal to the difference between B and C (this would require an interval scale), or that C has three times as much of the characteristic as A (this would require a ratio scale). Finally, the assumed ordinal scale also implies that the test scores should be analyzed by nonparametric statistical techniques. (For more on this issue, see Santor and Ramsay's article [1998] on item response theory in this issue.)

Another caveat concerns the use of physical measurements to stand for, or relate to, psychological values. Just because a measurement lies on a physical ratio scale (e.g., voltage or force), this by itself never implies a strength of scale for its psychological referent. For instance, although heart rate (beats per min) is measured on a ratio scale, initially it must be considered to be only monotonically related (at best) to the psychological variable of stress (see Tomarken, 1995, for a detailed discussion of such methodological and conceptual issues in psychophysiological assessment). In principle, such a scale might be elevated to a higher level, but only after meeting the required conditions of the stronger hoped-for scale (e.g., Krantz et al., 1971; Roberts, 1979). An apparent exception to this principle is when time is used in completely specified methodologies of mental architecture discovery (e.g., Schweickert, 1978; Sternberg, 1969; Townsend & Ashby, 1983; see Townsend, 1992, for a discussion of this and related issues). Used in this way, time is not representing a psychological thing, but merely is serving as a measuring instrument itself, much as the notion of probability does in theoretical models. Scientific progress is impeded when such strength of scale issues are disregarded.

4. Using deterministic theories and methods to model inherently probabilistic phenomena. Most psychologists are aware, at least vaguely, that probabilistic thinking has displaced most deterministic thinking in the hard sciences. Nevertheless, many of these psychologists continue to use theoretical constructs, sampling methods, measurement models, data-analytic strategies, and inference processes in their own work that are grounded in a deterministic framework. Obviously, there are occasions when much can be learned by simplifying a problem through the purposeful and judicious use of deterministic assumptions. But deterministic theories and methods continue to dominate psychological assessment even when they are not the most rational choice. This suggests that some psychologists may not be aware of the theoretical and practical limitations of deterministic approaches within psychology, and the potential advantages of more probabilistic alternatives. Perhaps this is because only limited coverage is given to formal probability theory in research design and statistical methods courses in doctoral programs in psychology (Gigerenzer et al., 1989).

The deterministic perspective is epitomized by the psychometric concept of the "true score" on psychological tests. According to this concept, an ideal number exists that represents accurately each participant's "true" scale value on the variable assessed by the test. This "true" value is reflected imperfectly by a person's observed test score, however, because of the contaminating influences of measurement error and other sources of "noise." But in theory, if an individual were tested repeatedly, the sample mean of these test scores would approach the person's "true score" asymptotically as the sample size increased. The "true score" concept not only reflects a reification of the test construct, but it also requires the assumption that the essence behind the "true score" is highly stable. This kind of reasoning has led to serious logical problems. Consider, for example, the concept of over-achiever, which is invoked when a person's task performance surpasses the level predicted on the basis of the person's "true score" on a test. This error of prediction is explained away post hoc by attributing it to a personality characteristic of the participant, rather than to weaknesses in the test, or to flaws in the "true score",concept and its assumptions. Of course, the "over-achiever" concept is an oxymoron. How can people possibly perform at levels beyond their performance capability?

The probabilistic perspective, in contrast, assumes that events are stochastic; that is, they are multiply determined by events that are inherently unpredictable themselves. When events are affected by random inputs, the result tends to be stochastic chaos (see Haynes, Blaine, & Meyer, 1995, and Heiby, 1995a, 1995b, for a discussion of deterministic chaos theory and psychological assessment). From this stochastic perspective, the goals of traditional deterministic assessment are illusory. Perhaps it is a reflection of our hubris as psychologists that we are so reluctant to abandon our fantasy of developing tests that can capture the essence of individuals in a single number, a number that allows us to predict unique and remote human events with precision. No amount of theoretical refinement or psychometric tinkering will give psychologists such a powerful crystal ball. The goals of probabilistic assessment are more realistic: to build dynamic, stochastic models with which to map the probability distributions for theoretically relevant phenomena, and to use these probability distributions to improve the accuracy with which we are able to estimate the likelihood of a range of events. In this respect, probabilistic assessment has more in common with meteorology than with astrology. Probabilistic scientific models, to the extent that they have utility, should help us (a) specify the probability distributions for our phenomena, (b) model quantitatively the functional relations between changing conditions and changes in these probability distributions, and (c) model the dynamic and stochastic processes that govern these conditional probabilities. Armed with such nomothetic knowledge about conditional probabilities, clinical scientists then should be in a better position to tackle more idiographic tasks, such as assessing the probability distributions of current and future events in the lives of individuals.

5. Using static measurement approaches to model inherently dynamic processes. Psychologists need to face this issue squarely. We rely almost exclusively on static measurement approaches. We do this not because such approaches are dictated by the inherent nature of our subject matter, but because we have not mastered the more challenging measurement approaches that are better suited to modeling the dynamic processes that interest us. Expediency sometimes may justify the use of static measures to test a simplified or preliminary model, but such occasions should be the exception, not the rule. If psychology is to advance as a scientific discipline, it simply must incorporate the previously neglected qualities of time and change as core elements at all levels of psychological models. Just as the grammatical rules of language require that a subject and verb must be consistent, our scientific models must follow basic structural rules: Dynamic theories require dynamic assessment methods!

The practice of talking dynamically while measuring statically can be seen clearly in the theories and methods of clinical psychologists. Clinical psychologists regularly offer verbal descriptions of putative dynamic relationships involving families, married couples, parents and children, and therapists and patients. With few exceptions, however, when these clinical investigators set out to test their verbal accounts empirically, they use strategies capable of yielding only static measurements (e.g., self-report questionnaires or interviews administered on one or two occasions). The problem is compounded when the resulting data, which typically lie on an ordinal scale, are analyzed by statistical methods that violate the theorems of axiomatic measurement (e.g., *F* or *t* tests).

On rare occasions, clinical investigators may use time-series analysis (e.g., Gottman, 1981) or path analysis (e.g., Patterson, Reid, & Dishion, 1992) to model the time and context-dependent changes in their participants. Unfortunately, many well-intentioned efforts to apply such tools tend to suffer from two problems: (a) the verbal dynamic theories do not make sufficiently precise and unique predictions; and (b) almost every experimental result could be accounted for with equal ease by several competing and seemingly contradictory theoretical models. In combination, these two problems make it virtually impossible to falsify the original verbal theory by such quantitative tests

These problems are illustrated by Mealey's (Mealey, 1995) sociobiological theory of the etiology of sociopathy. Critics of

Mealey's theory (McFall, Townsend, & Viken, 1995) have argued that such dynamic accounts must be specified with sufficient quantitative rigor that evidence from empirical tests will alter the standing of competing theoretical ideas, with some ideas being discarded outright and the remainder being assigned differential weights reflecting their probable truth value. Scientific assessment should yield information that leads to useful inferences. If theories are underspecified, or if they do not make unique predictions, or if their instrumental methods and measurement models are inappropriately matched, then psychological assessments simply cannot yield the kind of information required for meaningful scientific inferences and decisions.

Information and Inference

We believe that an information-focused approach to building and testing theoretical models is one of the most promising and powerful strategies currently available to psychologists, precisely because it permits strong scientific inferences. Ideally, scientific theories lead to risky, falsifiable predictions that go beyond the events that gave rise to them in the first place; they not only specify what is likely to occur, but also forbid other things from occurring; and they explicate the mechanisms, control structures, and system architectures governing the predicted events (Feynman, 1985; Popper, 1962). The more precise the theoretical predictions, assessment methods, measurement models, and data reduction and analysis strategies, the greater the risk of falsification (Meehl, 1978).

In the "soft science" (Meehl, 1978) of psychology, however, theories rarely, if ever, are tossed out on the basis of definitive experimental tests. One might question whether any of our current psychological theories even are capable of being falsified. Indeed, cynics might argue that no psychological theory ever has been discarded for any sound reason whatsoever. Psychological theories seldom fade away, let alone die.

What holds for theories also holds for their associated methods. For example, many clinical psychologists persist in using projective tests (e.g., Rorschach; Draw-a-House-Tree-Person) despite the lack of support for the parent theories and strong countervailing evidence against the methods (e.g., Chapman & Chapman, 1969; Wood, Nezworski, & Stejskal, 1996). No area of psychology seems free from the kudzu-like overgrowth of methods that refuse to die. In cognitive psychology, for instance, some experimenters persist in using inappropriate methods in an effort to discriminate between parallel (simultaneous) and serial (one-at-a-time) architectures in information-processing systems (e.g., methods based on the fallacious assumptions that all serial systems should predict increasing linear functions of load, whereas parallel processes should predict flat functions of load). Such methods live on even though they have been shown to be weak and ineffectual (e.g., Townsend, 1971b, Wolfe, 1998) and even though alternative methods capable of making such discriminations are available (e.g., Schweickert & Townsend, 1989; Townsend, 1990b; Townsend & Ashby, 1983).

Scientific theories and methods should not survive in the face of clear-cut negative evidence. Perhaps it is unrealistic to expect that the evidence from a single study will be sufficiently clear-cut to determine the fate of a given theory or method; however, it does seem reasonable to expect that the cumulative weight of empirical results gathered from multiple studies over time eventually should be sufficiently clear to have some theoretical import and practical implication.

Faced with the difficulty of conducting ideal definitive experiments, many psychologists believe that an information-focused approach provides a useful alternative for building, testing, and evaluating the implications of theoretical models. This is not a unified approach with a precisely defined set of methods; rather, it is a broad, general approach that has evolved over the years. It is a braided cord of conceptual and methodological strands that can be traced to several influential sources: (a) information theory (Shannon & Weaver, 1949); (b) signal-detection theory (e.g., Tanner & Swets, 1961, in psychology; W. W. Peterson & Birdsall, 1953, in electrical engineering); (c) cybernetics, or feedback control theory (e.g., R. Ashby, 1952; Wiener, 1948); (d) the theory of automata, especially in relation to the brain (e.g., von Neumann, 1958); (e) the theory of games, decisions, and utility (e.g., von Neumann & Morgenstern, 1944); and (f) artificial intelligence (e.g., Newell, Shaw, & Simon, 1958). Work in these divergent areas has shown psychologists how to study complex mental processes without confronting the philosophical conundrums that (along with introspectionism) drove early psychologists to behaviorism, psychophysics, and physiology.

The information-focused approach is dominant, for example, in contemporary cognitive science. In general, cognitive science views the organism as a set of processing subsystems, ranging from input to output, sometimes with feedback loops. The exact number and kind of subsystems depends on the particular psychological problem being studied, but they interact so as to fulfill the task characteristics prescribed by the experimenter (or by nature). Two examples of this approach are Newell's unified theory based on SOAR (Newell, 1990), a complex, computer-oriented, intelligent system; and Grossberg's neurally oriented connectionist model (Grossberg, 1987). Cognitive scientists regularly use information-focused methods to help them test their theories of information processing in humans.

The information-focused approach, especially as it has been developed and refined in cognitive science, offers a number of potential advantages to assessors. For example, signal detection theory (SDT; Green & Swets, 1974; Macmillan & Creelman, 1991) and its descendants, such as general recognition theory (GRT, F. G. Ashby & Townsend, 1986) and choice theory (CT; Luce, 1963), provide separate assessments of a system's sensitivity to signals, on the one hand, and its response biases (thresholds, criterion levels, decision boundaries, cutting scores), on the other hand. By providing separate estimates, they solve the thorny and long-standing problems ordinarily encountered when studying phenomena with extremely high or low base rates, problems that especially have been acute for clinical scientists (Meehl & Rosen, 1955). Furthermore, SDT, GRT, CT, and other related approaches offer several well-studied, content-free measurement models and statistical methods with which to assess quantitatively the amount of information provided by a system under different conditions (e.g., the receiver operating characteristic curve in SDT; see McFall & Treat, in press). These assessment tools come complete with a host of well-developed, thoroughly studied, content-free assessment tasks and paradigms that can be (and have been) adapted readily for the study of a wide variety of specific problems (e.g., see Swets, 1996, for a survey of the diverse range of SDT applications). As a bonus, these methods typically are designed to assess participants' optimal performance under dynamic conditions; they seldom require deception; and they usually can be administered and scored efficiently by computer, thereby promoting standardization. Finally, these assessment tasks typically are designed to sample participants' actual information-processing performance and to yield conditional probability distributions for performance data consistent with probabilistic, stochastic, dynamic models of behavior. Thus, they satisfy most of McClelland's (1973) six recommendations for ideal assessment (paraphrased, these are (a) use criterion sampling, (b) assess dynamic processes, (c) make evaluative criteria public and explicit, (d) focus on competencies, (e) focus on both operant and respondent behaviors, and (f) focus on actual cognitive processing).

Investigators who launch research programs into difficult and complex problems often start by using simplistic models and analytic methods. This can be a reasonable and practical strategy. Over time, however, as more is learned about these complex problems, more rigorous models and methods are required. The most rigorous exemplars of the information-focused approach, such as those mentioned above, take their form as mathematical models. Obviously, there can be nonsense, sloppiness, and ineptitude in modeling, just as in any human endeavor, but the process has built-in constraints that lower this likelihood substantially. The first step in such modeling is to translate one's verbal expectations and assumptions into an explicit quantitative expression. One then begins deriving predictions and other characteristics of the model, building a hierarchy of increasingly rigorous formulations. At the lowest level, researchers simply may test out their intuitions about the model's behavior through computer simulations designed to act out the model's hypothesized mechanisms and processes. At a higher level, the model's predictions are specified completely by formulas. Because such formulas are complicated, it sometimes is impossible to anticipate their behavior under all conditions; therefore, their behavior can be observed in an empirical sense, using a digital computer. Starting with the explicit formulas, particular parameter values are inserted into the computations and the resulting numerical outcomes are observed; these outcomes, in turn, are compared against the modeler's expectations.

The next and highest level of modeling, in our opinion, is to derive qualitative predictions of the model from the formulas without further numerical computation. In fact, sometimes it is possible to investigate whole classes of models, with each class representing distinct and contrasting psychological principles or hypotheses. Such investigations, done with care, can provide qualitative tests that are free of specific distributional assumptions and do not require the traditional parameter estimation and fitting routines. Townsend's (1984, 1990a) work on parallel versus serial processing illustrates this strategy. Examples of qualitative properties to be predicted are increases, decreases, or curvatures of data functions. Another example would be predicted inequalities among important theoretical entities. Yet another would be differential predictions of the patterns of stability and change across varying conditions. In some modeling approaches, in fact, only the qualitative form of the data may be predicted, with other system variables, such as time, left to be scaled properly at a later time. This type of qualitative approach is not always easy, and frankly, not everyone agrees that it is the preferred strategy (e.g., Van Zandt & Ratcliff, 1995).

Typically, to assess quantitatively how well models fit the data, investigators use such methods as chi-square, maximum likelihood, or least-square fit. One must be wary at this point. Testing the adequacy of a model's fit is based on exactly the reverse logic of the usual null hypothesis testing in psychology. Ordinarily, the hypothesis of interest (H_1) predicts a difference, the null hypothesis (H_0) predicts no difference, and a relatively low alpha value (.05) is chosen to minimize the probability of a Type I error. When model fitting, however, most investigators hope that their model's predictions do not differ significantly from their data; that is, they hope that the null hypothesis does not get rejected. In this case, using an alpha of .05 to reject the null hypothesis is inappropriate and misleading, as it stacks the deck in favor of finding no difference and of concluding that the model fits. Note that the null hypothesis also is less likely to be rejected when statistical power is low. Obviously, investigators must attend to these implications of the reverse logic when designing statistical tests of their model's fit.

Some investigators use a correlation measure (R^2) to assess the degree of correspondence between predicted and observed data points, but this approach is not very powerful; even data sets in which predicted and observed values deviate considerably can yield correlation coefficients of .95 or above. In any model test, the number of parameters generally should not exceed the number of degrees of freedom in the data. As a rule, when two models fit equally well, the one with the fewest parameters would be considered more powerful, although there are exceptions to this rule (e.g., Bamber & van Santen, 1985; Marley, 1992; Townsend & Landon, 1982).

Given the difficulty of evaluating the fit of a single model, a stronger strategy may be to compare the relative fit of two or more competitive models, particularly if these are based on countervailing psychological principles. Assuming that they have the same number of parameters, the explanation provided by the best fitting model is judged to be closest to the truth. Another common strategy is a comparison between a general model and a nested or restricted version of the general model (Wickens, 1982); essentially, one or more parameters of the general model are omitted from the restricted version, and their respective fits to the data are compared to determine what additional information, if any, is provided by the general model, relative to the information provided by the restricted model.

Competitive model testing certainly is one of the best ways to move toward theoretical truth. It is not without its limitations, however. Some models simply seem to have a broader ability to handle data. For instance, the overlap model (Townsend, 1971a, 1971c) of pattern recognition cannot predict certain patterns of data that can be predicted by the similarity choice model (a version of Luce's choice model [Luce, 1963; Shepard, 1958]), even when the two models have the same number of parameters (Pachella, Smith, & Stanowich, 1978; Townsend, 1971a, 1971c). Comparing models with different scopes can be like comparing pork and beef.

Recent advances in model testing strategies now make it possible to assess simultaneously not only a model's fit, but also its complexity (Myung & Pitt, 1997). For models with a given number of parameters, higher complexity implies an ability to handle more diverse data. Higher complexity also may imply a more lax set of qualitative predictions about the data, although this remains to be demonstrated empirically. This approach to model testing also carries its own assumptions, but in our opinion it is one of the more promising developments to come along in recent years.

Model building and testing is not a one-shot affair, of course, but is an iterative, boot-strapping process that unfolds over time through a series of smaller-scale conjecture and refutation stages—a process that Platt (1964) has called "strong inference." The process of building a general model, from this perspective, consists of a series of choices as one climbs the branching limbs of a decision tree. At each fork in the tree, one must decide which branch to follow. These decisions are guided by the results of empirical tests comparing the relative strengths of the plausible options at each juncture.[4] The strong inference approach is illustrated by the work of Massaro (e.g., 1989), who used his fuzzy logical model of processing to investigate a panoply of questions concerning visual and acoustic information processing.

In psychological research, this decision process rarely is clear-cut or certain; we seldom know all of our options or have access to all of the relevant evidence. Despite the uncertainty, we still must choose. Once we have chosen, we then must attend to the consequences of our choices. We often can learn as much from "wrong turns" as from lucky guesses. Thus, scientific model building is a discovery process: We simply cannot know what we will find until we get there.

Assessment in Clinical Science

We have asserted that good scientific models have utility. They enable us to describe, explain, detect, and predict events with greater accuracy than without them. The benefits of good scientific models are evident all around us, from the food we eat and clothes we wear, to the tools and appliances we use, to the medicine that keeps us healthy, and to the energy that keeps things running. In psychology, good models offer many potential benefits as well. Arguably their greatest potential impact would be on the interpersonal and intrapersonal human problems that have been the primary focus of clinical psychologists. It is a paradox, therefore, that clinical psychologists, who may have the most to gain, have not played a more active and influential role in the development and testing of rigorous quantitative models in psychology.

Perhaps clinical psychologists have been distracted from full participation by practical and professional preoccupations. Indeed, Woodworth warned against this possibility even as the subspecialty of clinical psychology was being born (Woodworth, 1937). His concerns seem to have been realized (Sechrest, 1992). The professionalization of clinical psychology has led, in part, to what Cronbach called the two worlds of psychology (Cronbach, 1957). Most clinical psychologists live in an applied world of psychotherapy and psychological testing and prediction, a world focused on providing idiographic solutions to the unique problems of individuals, couples, or families. Most other psychologists, in contrast, live in an abstract world of model building and testing aimed at finding nomothetic solutions to a wide range of general problems.

As we noted earlier, nomothetic knowledge is a prerequisite to valid idiographic solutions, so one might expect that clinical psychologists would attend closely to developments in the nomothetic world. This is not the case, however. It is as if the two worlds of psychology not only moved in independent orbits, but also occupied different corners of the universe. Many clinical psychologists seem unaware of and unaffected by developments and discoveries in the nomothetic world, even when these potentially could help them understand, assess, treat, predict, and prevent psychopathology and other suffering.

Perhaps nowhere is the gulf between the two worlds of psychology more evident than in the area of cognitive theory and assessment. In both worlds, cognition is an important inferred cause of observable behavior; the similarity stops there, however. The two sides may use similar language at times, giving the illusion of some connection, but a closer look at the constructs and methods behind their common language belies the surface similarities. The fundamental differences are revealed by comparisons between the concepts and methods of cognitive–behavioral clinicians, on the one hand, and the concepts and methods of contemporary cognitive scientists, on the other hand (McFall, Treat, & Viken, 1997, 1998).

The prototypic cognitive–behavioral clinician merely assumes causality, without a quantitative theoretical model or empirical evidence; whatever behavior is observed must have resulted from a person's cognitions.[5] This explanation is similar to a homunculus account, in which all behavior is attributed post hoc to some thing (cognition; homunculus) in the head; but the postulated causal agent is not defined precisely (Barlow, 1996). This merely pushes the explanation back one step, and it begs the question of how the putative agent actually causes the behavior. Because the explanation generates no risky predictions, it is unfalsifiable. Because it explains everything (post hoc), it explains nothing.

Linguistically, clinicians typically represent cognitions either as nouns (i.e., as things that people have, that impinge on people, or with which people struggle) or as verbs or verb–adjective combinations (i.e., either as mental experiences or as experiential affective states). In both cases, these are phenomenological, subjective, and often transient. Examples in the noun category are attitudes, attributions, beliefs, goals, and thoughts. Examples of the verb form are to attribute, to expect, and to think. In the verb–adjective form, the verb "to feel" is combined with affective states, for example, anxious, depressed, or stressed.

[4] Newell (1973) has been critical of "playing 20 questions with nature." However, the strong inference approach differs from Newell's "game," in that strong inference simply represents a logical scheme for systematically and sequentially attacking the vast array of theoretical questions one inevitably must address when building a model. Some questions (e.g., the parallel vs. serial processing question) are so central, in fact, that work on a model is stymied until they are resolved.

[5] We acknowledge the dangers of such generalizations, and we expect that some will dismiss our description as a caricature or "straw person." However, we believe that the overall image in our broadbrush portrait would not change fundamentally if it had been rendered in a pointillist style

The cognitive–behavioral clinician considers cognitions to be subjective and internal phenomena that cannot be observed directly by the clinician. Therefore, to assess a person's cognitions, the clinician relies primarily on indirect, introspective, self-report methods, such as paper-and-pencil questionnaires or clinical interviews. When it comes to treatment, most cognitive–behavioral therapies are designed either to counter and restructure patients' irrational cognitions through verbal reasoning or to eliminate disruptive cognitions through exposure-based extinction procedures. The strongest empirical support for the clinician's conception of cognition comes from therapy outcome studies in which patients with disabling problems, such as anxiety disorders, depression, or eating disorders, report more improvement on self-report measures following cognitive–behavioral therapy than following either placebo-control treatment or no treatment. It would be a mistake, of course, to infer from these studies either that the therapeutic gains were due to cognitive changes (which cannot be observed directly), or that the therapeutic gains establish the validity of the theory behind cognitive–behavioral treatment (as opposed to all the other plausible explanations for the effects). Both inferences would be examples of the logical fallacy "affirming the *consequent*" (e.g., UFOs prove that Martians are invading our planet; decreased pain following acupuncture proves the ancient Chinese theory of bodily energy).

The clinician's views on cognition contrast sharply with those of the contemporary cognitive scientist.[6] To the cognitive scientist, cognition is not a static thing or a subjective experience, but a process. Specifically, *cognition* is a general label for the complex, dynamic processes "in the black box," by which humans transform stimulus information, on the input side, into observable actions, on the output side. The aim of cognitive science is to build and test general theoretical models that improve our ability to describe, explain, and predict the operations of this human information-processing system. Examples of cognitive processing operations are categorization, classification, feature detection, recall, and recognition. Whereas cognitive processing "in the black box" cannot be observed directly, neither do human beings have access to, or accurate knowledge of, their own cognitive processing. Therefore, falsifiable quantitative predictions derived from theoretical models of these processes must be assessed systematically through direct samples, obtained under controlled conditions, of observable performance on tasks that tap such operations as categorization, classification, and feature detection. These tasks simply require participants to categorize, classify, and detect. Introspective methods (which were rejected by most cognitive scientists long ago) seldom are used. Participants' self-reports about internal processes, if used at all, are treated no differently than other output responses from the system. That is, they are dependent variables used to test theoretical predictions; the content of self-reports is given no special status or validity.

As these comparisons reveal, clinical psychologists simply have not kept abreast of advances in cognitive science over the last half century. It is time for this to change. Clinical science should become integrated with cognitive science—at all eight levels of the multilayered model outlined in this article. The integration must start at the top, of course, with a critical reexamination of the cognitive theoretical constructs (and their referents) that have gained such wide acceptance in clinical psychology, but are so discrepant from those in cognitive science. Attention to instrumentation and measurement issues would follow logically from attention to these theoretical concerns. The first step might be to reevaluate critically the popular clinical theories for specific problems, such as anxiety disorders or depression. We should look skeptically, for example, at the widely held view that depression is the product of maladaptive cognitions as things people have in their heads (e.g., depressive attributions, negative self-schemas, or low self-esteem).

Are such clinical theories well conceived and empirically supported, or are they tautological? Should we worry, for example, when the dependent variable (depression) and the independent variables (hypothesized causes, such as attributions, schemas, or self-esteem) all are assessed by using highly similar methods (introspective self-reports) to get information from a single source (the participant) at one point in time? To the extent that these methods yield convergent results, is this evidence of construct validity, or is it merely a reflection of method reliability? When people who report feeling depressed on questionnaires also endorse negative self-relevant items on questionnaires at the same time, is this really surprising? Does it really tell us anything about cause–effect relationships? Do participants' introspective reports provide veridical accounts of their cognitive processes? Where in this system are the counterintuitive, risky, falsifiable predictions that are the hallmark of a good scientific model?

The integration process cannot bypass theoretical issues, skipping directly to the method layer. To do this would be to fall prey to the logical errors we discussed earlier (see *Logical Structure of Assessment* section). For example, methods cannot stand alone, they are not atheoretical or assumption free, and they cannot serve as automatic inference machines. Nevertheless, some clinical investigators have treated published cognitive assessment methods (e.g., Wisconsin Card Sort Test, Stroop Test, and Dichotic Listening Test) as though they were off-the-shelf standardized tests, like the Stanford–Binet. (Even the Stanford–Binet may be questionable, when used to represent a reification of intelligence.) Experimental cognitive tasks might be legitimate choices in some clinical studies, but only if their underlying constructs and measurement models actually are congruent with the investigator's theory and measurement model (Levin, Yurgelun-Todd, & Craft, 1989). It is an error to assume (a) that a published assessment tool, such as the Wisconsin Card Sort Test (Heaton, 1981), is a valid test of a well-established construct (e.g., executive function); (b) that the tool can be taken out of its theoretical context and used with no excess theoretical baggage; (c) that the tool can be transplanted to a different problem, paradigm, or population without risk or loss; and (d) that empirical support for the tool, gathered under different conditions, travels with the tool, regardless of how it is used.

How might things change if clinical scientists were to adopt the theoretical perspective of contemporary cognitive science? For starters, they would build theoretical models of the processes by which humans transform stimulus information into action.

[6] Here we go again with caricatures! But, as with a "magic eye" picture, try to look through the quibbles to see the embedded contrasts.

and this change in theoretical focus would have ramifications for all other layers of their models. Their key constructs would focus on the referents for these dynamic, probabilistic processes and their characteristics. Their instrumental methods for sampling these referents would not rely on introspective self-reports, but would emphasize systematic observations of actual performance on tasks designed to sample the specific processes of interest. Their measurement models would be based on the strongest possible measurement scales, would be rigorously quantitative, and would provide critical tests of risky predictions derived from the theoretical models.

Although too many clinical psychologists still use cognitive concepts and methods that are at variance with those in contemporary cognitive science, there is reason for optimism, nonetheless. Some clinical scientists have been building bridges across the gulf, adapting the tools of cognitive science to study clinically relevant phenomena. We cannot possibly list all of these pioneers, nor could we do justice to their work by trying to summarize it in the remaining space. Perhaps we can give readers a small taste of this integrative research, however, by citing a few examples. One needs to see only one unicorn to know that they exist. If skeptical readers pursue the leads we highlight on their own, perhaps they will be convinced that clinical and cognitive science can be integrated. These examples also may serve as concrete models for investigators who find the idea *appealing in the abstract, but do not know how to get started* in this new direction.

One of the most prolific, persistent, and influential clinical-cognitive scientists is Peter J. Lang. He has devoted his distinguished career to building and testing a theoretical model of emotion, with a special focus on fear and anxiety (e.g., Lang, 1995; Lang & Cuthbert, 1984). His evolving model of emotion has been formulated explicitly within an information-processing framework, with firm anchors in the underlying neurophysiology. He construes emotions as response dispositions, as states of vigilant readiness to respond, that are driven by two motivational systems, appetitive and aversive. Arousal, in this model, represents the level of metabolic and neural activation of the appetitive and aversive systems. Dispositions are centrally activated systemic responses that mobilize the organism in preparation for such action. Emotions, in this view, occur when these highly motivated action potentials are delayed or inhibited. Referents for these affective states of readiness in humans can be found in three response systems: verbal behavior, somatic and autonomic behavior, and skeletal-muscular behavior.

To test this model of emotion, Lang and his colleagues have relied heavily on psychophysiological methods to sample specific referents (e.g., heart-rate acceleration, blood-pressure level, electrodermal reactivity, corrugator and zygomatic EMG response) for the theory's key constructs.[7] For example, they have assessed participants' *startle responses (EMG amplitude and* latency of the early eyeblink reflex) to acoustic (50-ms burst of 95-db white noise) and visual (strobe-light flash) probes presented under different emotional conditions (i.e., while participants viewed photos differing in their normatively rated affective valence and level of arousal). Using this assessment paradigm, they have tested the theoretical proposition that the amplitude of participants' startle responses is a function of participants' current, active disposition (appetitive or aversive). *Specifically, when participants are processing appetitive information*, they should show a diminished startle response, relative to when they are processing aversive information. In general, the data support these expectations.

Lang's model not only is coherently integrated and compelling, but it also has led to risky predictions while avoiding the problems noted above (circularity, common method variance, introspective assessment of cognitive processes). Rather than following an off-the-shelf approach to assessment, innovative assessment methods have been developed to fit specific theoretical questions. Finally, the information yielded by this research has important implications for our understanding of emotion, generally, and for the assessment, prediction, and treatment of such clinical problems as fear and anxiety.

Richard W. J. Neufeld (editor of this Special Section) is another investigator whose work has exemplified a persistent commitment to the integration of clinical, cognitive, and neural science. Specifically, Neufeld and his colleagues have focused on building and testing neuropsychological theories of schizophrenia, with particular emphasis on explicating the different neurophysiological architectures that underlie specific subtypes of schizophrenia, such as paranoid versus nonparanoid (e.g., Highgate-Maynard & Neufeld, 1986). Searching for differential deficits in schizophrenia is nothing new in psychology; however, Neufeld's approach differs fundamentally from the norm. *Whereas many others have pursued an atheoretical "power"* strategy, looking for group differences anywhere and everywhere they could find them, Neufeld has pursued a theory-driven, quantitatively rigorous strategy, grounded in current knowledge regarding cognitive processes and neural systems.

Neufeld and his colleagues have adopted an impressive array of theoretical constructs; experimental tasks and paradigms; and measurement models and analytic tools from cognitive science and neuroscience in their efforts to model the neural architectures underlying different patterns of symptomatic behavior in schizophrenia (see Neufeld & Williamson, 1996, for a review). We will cite only a few examples here to illustrate the scope of their work: (a) They have used multidimensional scaling techniques to compare perceptions of verbal information in individuals with and without schizophrenia; (b) they have used template matching in memory search tasks to compare encoding of information in paranoid and nonparanoid subtypes of schizophrenia; (c) they have examined mnemonic organization and recall of categorical word lists in persons with schizophrenia; (d) they have assessed lateralization abnormalities in a visual recognition task where items were presented tachistoscopically to alternate visual fields under varying levels of demand; and (e) they have looked at specific temporal features of performance, such as mean response latencies and variance in response times (across trials, within subjects who differed in symptoms), on tasks in which the encoding demand, or load, was varied systematically. In the latter work, for example, rigorous mathematical models representing competing theoretical hypotheses were fit to the available data from people with para

[7] Indeed, many of Lang's former students (e.g., Michael Kozak, Barbara Melamed, Gregory Miller, Robert Simons, Scott Vrana) now are independent investigators making influential clinical-cognitive contributions of their own.

noid schizophrenia and controls. Results indicated that the groups differed in the number, or efficiency, of their covert encoding subprocesses, but not in their overall rate of processing. Across studies, Neufeld and his colleagues consistently have sought to gather evidence that can inform our evolving neurophysiological models of psychopathology. Although Neufeld's primary focus has been on schizophrenia, he has been a strong advocate for the integration of clinical, cognitive, and neural science across the spectrum of psychological disorders (Neufeld, 1995).

We have identified two investigators whose work exemplifies the kind of integrative and rigorous approach to clinical science that we have been advocating in this article. Although we might have cited other examples, as well, our purpose here is not to provide an exhaustive list, but merely to provide concrete evidence that the abstract ideal can be achieved. We certainly have been pursuing this ideal in our own research on clinical problems. One or both of us have collaborated with varying subsets of colleagues from clinical science (e.g., Richard J. Viken, Teresa A. Treat, A. Michele Lease, Shoma S. Ghose), cognitive science (e.g., Robert M. Nosofsky, John K. Kruschke, David B. MacKay), and neuroscience (e.g., Joseph E. Steinmetz, Donald Katz, Jo Ann Tracy) on a variety of projects aimed at exploring four clinically relevant problems (e.g., sexual coercion by men, eating disorders among women, children's peer groups, obsessive-compulsive disorder) using a variety of tools adopted from cognitive science and neuroscience (MDS, SDT, CT, and mathematical modeling techniques; similarity ratings, prototype classification, category learning, visual search, recognition memory, and classical eye-blink conditioning experimental tasks and paradigms).

To date, the results of these hybrid collaborations have been encouraging. For example, we used MDS to map the perceptual organization men imposed on a set of photos of women, and then we used the attention weights captured by these perceptual maps to predict the men's performance in category learning tasks. As we expected, the men performed best when their perceptual map was congruent with the category structure of the learning task. We also used MDS to map women's perceptions of photos of other women. These maps revealed that women who admitted to bulimic behaviors also showed greater attention than controls to the body-size dimension in the photos, and less attention to the affect dimension. In studies such as these, we have begun to establish theoretically meaningful links between specific types of clinical symptoms, on the one hand, and patterns of information processing, assessed through methods adapted from cognitive and neuroscience, on the other hand (see McFall et al., 1998, for a summary of initial results from these and other research projects).

In fairness, we cannot conclude without offering a few caveats. We have attempted to show that there is reason for optimism about the future of psychological assessment in clinical science. However, we hope that in the process, we have not left false impressions (a) that the required changes will be easy, or (b) that all other problems will disappear if these changes are made. Indeed, we can identify several significant, unresolved issues in psychological assessment that require further attention.

First, traditional approaches to statistical inference were not designed for the kinds of idiographic assessment and prediction tasks that traditionally have been the focus of clinical psychologists (Gigerenzer & Murray, 1987). To begin with, given the low base rates of most clinical problems (typically below 5%), predictions based on clinical assessments almost never can surpass the overall accuracy of base-rate predictions. To make matters worse, however, psychologists working in clinical settings typically are not attempting to make general probabilistic predictions for a population, but are attempting to make inverse probability predictions for individual cases. For example, these clinicians usually are not concerned with estimating the likelihood that persons with schizophrenia in general will display certain characteristics; rather, they usually are concerned with estimating the likelihood that a particular individual, who is displaying certain characteristics, is suffering from schizophrenia. The probability of a characteristic, C, given a diagnosis, D, *is not* necessarily the same as the probability of the diagnosis given the characteristic; that is, $(C|D)$ and $(D|C)$ are not equivalent. In combination, the related problems of low base rates and inverse probability predictions make it extremely difficult to achieve incremental validity in clinical predictions of individual cases (e.g., clinical case diagnosis, child custody disputes, and parole board decisions).

Second, the approach to psychological assessment that we have advocated in this article is time and labor intensive, of necessity. There seldom are ready-made solutions to the assessment problems that arise in connection with cutting-edge theoretical questions. This means that good psychological assessment often necessitates the handcrafting of tailor-made solutions. This implies, in turn, that psychological assessment, at least in most contexts, cannot be treated in an automated, off-the-shelf manner. Clinical scientists seldom can function as technicians, relying on formula approaches; instead, they must have the knowledge and skills to devise creative solutions to specific assessment problems.

If the aim of assessment is to reduce uncertainty, as we asserted at the outset, then clinical scientists necessarily are pursuing a moving target. As new information is acquired, new uncertainties emerge, raising new questions that require new assessment strategies. Psychologists capable of "thinking outside the box" will be in a better position to develop effective solutions to these evolving assessment needs. Also, psychologists are more likely to show such creative thinking if they are aware of developments outside of their own narrow specialty, and they are willing to explore, borrow, and adapt promising concepts and methods wherever they can find them. It is in this spirit, then, that we have argued in this article that an integration of clinical science with cognitive science and neuroscience should lead to significant advances in psychological assessment.

References

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93,* 154–179.

Ashby, R. (1952). *Design for a brain.* New York: Wiley.

Astin, A. W. (1961). The functional autonomy of psychotherapy. *American Psychologist, 16,* 75–78.

Bamber, D., & van Santen, J. P. H. (1985). How many parameters can a model have and still be testable? *Journal of Mathematical Psychology, 29,* 443–473.

Barlow, H. (1996) Banishing the homunculus. In D. C. Knill & W

Richards (Eds.), *Perception as Bayesian inference* (pp. 425–450). Cambridge, England: Cambridge University Press.

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

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 271–287.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

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

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674.

Feynman, R. P. (1985). *Surely you're joking, Mr. Feynman!* New York: Norton.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286–299.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life.* Cambridge, England: Cambridge University Press.

Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists.* Cambridge, England: Cambridge University Press.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics* (Rev. ed.). Huntington, NY: R. F. Krieger.

Grossberg, S. (1987). *The adaptive brain, Vol. 1.* Amsterdam: North Holland.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323.

Harris, R. J. (Ed.). (1997). Special section: Ban the significance test? *Psychological Science, 8,* 1–20.

Haynes, S. N., Blaine, D., & Meyer, K. (1995). Dynamic models for psychological assessment: Phase space functions. *Psychological Assessment, 7,* 17–24.

Heaton, R. (1981). *Wisconsin Card Sorting Manual.* Odessa, FL: Psychological Assessment Resources.

Heiby, E. M. (1995a). Assessment of behavioral chaos with a focus on transitions in depression. *Psychological Assessment, 7,* 10–16.

Heiby, E. M. (1995b). Chaos theory, nonlinear dynamical models, and psychological assessment. *Psychological Assessment, 7,* 5–9.

Highgate-Maynard, S., & Neufeld, R. W. J. (1986). Schizophrenic memory-search performance involving nonverbal stimulus properties. *Journal of Abnormal Psychology, 95,* 67–73.

Kantowitz, G. H., & Knight, J. L., Jr. (1976). Testing tapping timesharing, II: Auditory secondary task. *Acta Psychologica, 40,* 343–362.

Kelly, G. A. (1955). *The psychology of personal constructs* (Vols. 1, 2). New York: Norton.

Kozak, M. J., & Miller, G. A. (1982). Hypothetical constructs versus intervening variable: A reappraisal of the three-systems model of anxiety assessment. *Behavioral Assessment, 4,* 347–358.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement,* Vol. 1. New York: Academic Press.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10,* 1–11.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Lang, P. J. (1968). Fear reduction and fear behavior: Problems in treating a construct. In J. M. Schlein (Ed.), *Research in psychotherapy, Vol. III* (pp. 91–102). Washington, DC: American Psychological Association.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist, 50,* 372–385.

Lang, P. J., & Cuthbert, B. N. (1984). Affective information processing and the assessment of anxiety. *Journal of Behavioral Assessment, 6,* 369–395.

Levin, S., Yurgelun-Todd, D., & Craft, S. (1989). Contributions of clinical neuropsychology to the study of schizophrenia. *Journal of Abnormal Psychology, 98,* 341–356.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, Vol. 1* (pp. 103–190). New York: Wiley.

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55,* 95–107.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* Cambridge, England: Cambridge University Press.

Marley, A. A. J. (1992). Developing and characterizing multidimensional Thurstone and Luce models for identification and preference. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 299–333). Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1989). Speech perception by ear and eye: A paradigm for psychological inquiry. *Behavioral and Brain Sciences, 12,* 741–794.

Matarazzo, J. D. (1990). Psychological assessment versus psychological testing: Validation from Binet to the school, clinic, and courtroom. *American Psychologist, 45,* 999–1017.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence." *American Psychologist, 28,* 1–14.

McFall, R. (1993). The essential role of theory in psychological assessment. In R. L. Glueckauf, L. B. Sechrest, G. R. Bond, & E. C. McDonel (Eds.), *Improving assessment in rehabilitation and health* (pp. 11–32). Newbury Park, CA: Sage.

McFall, R. M. (1996). Making psychology incorruptible. *Applied & Preventive Psychology, 5,* 9–15.

McFall, R. M., Townsend, J. T., & Viken, R. J. (1995). Diathesis-stress model or "just so" story? *Behavioral and Brain Sciences, 18,* 565–566.

McFall, R. M., & Treat, T. A. (in press). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology.*

McFall, R. M., Treat, T. A., & Viken, R. J. (1997). Contributions of cognitive theory to new behavioral treatments. *Psychological Science, 8,* 174–176.

McFall, R. M., Treat, T. A., & Viken, R. J. (1998). Contemporary cognitive approaches to studying clinical problems. In D. K. Routh & R. J. DeRubeis (Eds.), *The science of clinical psychology* (pp. 163–197). Washington, DC: American Psychological Association.

Mealey, L. (1995). The sociobiology of sociopathy: An integrated evolutionary model. *Behavioral and Brain Sciences, 18,* 523–599.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1971). High school yearbooks: A reply to Schwarz. *Journal of Abnormal Psychology, 77,* 143–148.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

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

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review, 4,* 79–95.

Neufeld, R. W. J. (1995). Formal touchstones of abnormal personality theory. *Behavioral and Brain Sciences, 18,* 567–568.

Neufeld, R. W. J., & Williamson, P. C. (1996). Neuropsychological correlates of positive symptoms: Delusions and hallucinations. In C. Pantelis, H. E. Nelson, & T. R. E. Barnes (Eds.), *Schizophrenia: A neuropsychological perspective* (pp. 205–235). New York: Wiley.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers in this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York: Academic Press.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review, 65,* 151–166.

Pachella, R. G., Smith, J. E. K., & Stanowich, K. E. (1978). Qualitative error analysis and speeded classification. In N. J. Castellan, Jr., & F. Restle (Eds.), *Cognitive theory. Vol. 3* (pp. 169–198). Hillsdale, NJ: Erlbaum.

Patterson, G. R., Reid, J. B., & Dishion, T. J. (1992). *Antisocial boys.* Eugene, OR: Castalia Publishing Company.

Peterson, D. R. (1996). Making psychology indispensable. *Applied & Preventive Psychology, 5,* 1–8.

Peterson, W. W., & Birdsall, T. G. (1953). *The theory of signal detectability* (Tech. Rep. No. 13). Ann Arbor: University of Michigan, Electronic Defense Group.

Platt, J. R. (1964, October). Strong inference. *Science, 146,* 347–353.

Polya, G. (1957). *How to solve it: A new aspect of mathematical method* (2nd ed.). Princeton, NJ: Princeton University Press.

Popper, K. (1962). *Conjectures & refutations.* New York: Basic Books.

Roberts, F. S. (1979). *Measurement theory with applications to decision making, utility, and the social sciences.* Reading, MA: Addison-Wesley.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 10,* 345–359.

Schweickert, R. (1978). A critical path generalization of the additive factor method: Analysis of the Stroop task. *Journal of Mathematical Psychology, 18,* 105–139.

Schweickert, R., & Townsend, J. T. (1989). A trichotomy method: Interactions of factors prolonging sequential and concurrent mental processes in stochastic PERT networks. *Journal of Mathematical Psychology, 33,* 328–347.

Sechrest, L. (1992). The past future of clinical psychology: A reflection on Woodworth (1937). *Journal of Consulting and Clinical Psychology, 60,* 18–23.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana, IL: University of Illinois Press.

Shepard, R. N. (1958). Stimulus and response generalization: Deduction of the generalization quotient from a trace model. *Psychological Review, 65,* 242–256.

Smith, J. M. (1984, November). Science and myth. *Natural History, 11,* 11–24.

Sternberg, S. (1969). Memory scanning: Mental processes revealed by reaction time experiments. *American Scientist, 57,* 421–457.

Stevens, S. S. (1946, June). On the theory of scales of measurement. *Science, 103,* 677–680.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychological diagnostics: Collected papers.* Hillsdale, NJ: Erlbaum.

Tanner, W. P., & Swets, J. A. (1961). A decision-making theory of visual detection. *Psychological Review, 61,* 401–409.

Tomarken, A. J. (1995). A psychometric perspective on psychophysiological measures. *Psychological Assessment, 7,* 387–395.

Townsend, J. T. (1971a). Alphabetic confusion: A test of models for individuals. *Perception & Psychophysics, 10,* 449–454.

Townsend, J. T. (1971b). A note on the identifiability of parallel and serial processes. *Perception & Psychophysics, 10,* 161–163.

Townsend, J. T. (1971c). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics, 9,* 40–50.

Townsend, J. T. (1975). The mind-body equation revisited. In Chung-Ying Cheng (Ed.), *Psychological problems in philosophy* (pp. 200–218). Honolulu: University of Hawaii Press.

Townsend, J. T. (1984). Uncovering mental processes with factorial experiments. *Journal of Mathematical Psychology, 28,* 363–400.

Townsend, J. T. (1990a). Serial versus parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science, 1,* 46–54.

Townsend, J. T. (1990b). Truth and consequences of ordinal differences in statistical distributions: Toward a theory of hierarchical inference. *Psychological Bulletin, 108,* 551–567.

Townsend, J. T. (1992). On the proper scale for reaction time. In H. Geissler, S. Link, & J. T. Townsend (Eds.), *Cognition, information processing, and psychophysics: Basic issues* (pp. 105–120). Hillsdale, NJ: Erlbaum.

Townsend, J. T. (1994). Methodology and statistics in the behavioral sciences: The old and the new. *Psychological Science, 5,* 321–325.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes.* Cambridge, England: Cambridge University Press.

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96,* 394–401.

Townsend, J. T., & Landon, D. E. (1982). An experimental and theoretical investigation of the constant ratio rule and other models of visual letter recognition. *Journal of Mathematical Psychology, 25,* 119–163.

Turner, M. B. (1967). *Philosophy and the science of behavior.* New York: Appleton-Century-Crofts.

Van Zandt, T., & Ratcliff, R. (1995). Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin and Review, 2,* 20–54.

von Neumann, J. (1958). *The computer and the brain.* New Haven, CT: Yale University Press.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

Wickens, T. D. (1982). *Models for behavior.* San Francisco: Freeman.

Wiener, N. (1948). *Cybernetics.* New York: Wiley.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science, 9,* 33–39.

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7,* 3–10.

Woodworth, R. S. (1937). The future of clinical psychology. *Journal of Consulting Psychology, 1,* 4–5.

Received February 20, 1998
Revision received June 5, 1998
Accepted August 5, 1998 ■