When Clinical Description Becomes Statistical Prediction

Drew Westen *Emory University*
Joel Weinberger *Adelphi University*

This article reconsiders the issue of clinical versus statistical prediction. The term clinical *is widely used to denote 1 pole of 2 independent axes: the observer whose data are being aggregated (clinician/expert vs. lay) and the method of aggregating those data (impressionistic vs. statistical). Fifty years of research suggests that when formulas are available, statistical aggregation outperforms informal, subjective aggregation much of the time. However, these data have little bearing on the question of whether, or under what conditions,* clinicians *can make reliable and valid observations and inferences at a level of generality relevant to practice or useful as data to be aggregated statistically. An emerging body of research suggests that clinical observations, just like lay observations, can be quantified using standard psychometric procedures, so that clinical description becomes statistical prediction.*

> The style and sequence of the [book] reflect my own ambivalence and real puzzlement, and I have deliberately left the document in this discursive form to retain the flavor of the mental conflict that besets most of us who do clinical work but try to be scientists. (Meehl, 1954, p. vi)

In 1954, Paul Meehl published his classic book on *Clinical Versus Statistical Prediction*. Clinical prediction referred to the use of an individual (an expert; in psychology, a clinician) to predict an event. Statistical prediction referred to the use of an actuarial formula to predict the same event. In the prototypical study reviewed by Meehl, the clinical expert had access to all of the information used to create the competing formula (and sometimes additional data). The clinician could combine the information in any way he or she saw fit, making use of clinical skill, intuition, and theoretical knowledge. In contrast, the mathematical equation had no flexibility.

In the vast majority of cases, the formula turned out to be at least as good a predictor as the clinician. Meehl's understanding of this finding was that the clinician combined the variables in an idiosyncratic manner, whereas the formula combined them in the way that past history had shown to be most predictive. In statistical terms, the clinician was an imperfect, unreliable generator of regression weights (see Goldberg, 1991).

Meehl's book touched off a decades-long debate about the reliability and validity of clinical judgment. The "hard" scientists savored the victory of statistics over clinical intuition; the "soft" psychologists railed against the devaluation of clinical expertise. The terms of the debate (and the attendant affect) seem little different today. Although psychologists have revisited the question of clinical versus statistical prediction many times since Meehl's book (e.g., Dawes, Faust, & Meehl, 1989; Holt, 1958; Sarbin, 1962; Sawyer, 1966), the weight of the evidence remains the same as it was in 1954: In the vast majority of studies, a good formula matches or trumps an intuitive clinical soothsayer (Grove, Zald, Lebow, Snitz, & Nelson, 2000).

In framing the clinical–statistical debate, Meehl (1954) used the term *clinical* to refer to a method of aggregating data (informal, unstructured vs. statistical, actuarial). We believe, however, that the debate since Meehl has often confounded the method of aggregation (unstructured judgment vs. statistical aggregation using algorithms refined over successive iterations) with the nature of the observer (clinician–expert vs. lay). Meehl was clear in defining clinical as a mode of data aggregation (and his collaborators have largely adhered to that definition; e.g., Dawes et al., 1989; Grove & Meehl, 1996; Grove et al., 2000). However, in broader psychological discourse, clinical has come to be used more broadly (and in accord with its standard English definition) to denote the judgments, inferences, observations, and practices of clinicians. The confusion of these two meanings of clinical has led to a widespread belief that empirical data have shown that the observations, thought processes, and beliefs of *clinicians* are seriously flawed (e.g., Tavris, 2003).

Consider the following excerpt from Meehl's obituary, published in the *APS Observer*: "Meehl's reputation spread with his 1954 book . . . in which he showed that statistical formulas were better than, or at least equal to, *clinicians* at predicting such things as what sort of treatment would best benefit a mentally ill person" (American Psychological Society Observer, 2003, p. 13; emphasis added). This statement is particularly problematic given

Preparation of this article was supported in part by National Institute of Mental Health Grants MH62377 and MH62378 to Drew Westen. We thank William Grove, Scott Lilienfeld, Keith Rayner, and George Stricker for their comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Drew Westen, Departments of Psychology and Psychiatry, Emory University, 532 North Kilgo Circle, Atlanta, GA 30322 or to Joel Weinberger, Derner Institute, Adelphi University, Box 701, Garden City, NY 11530. E-mail: dwesten@emory.edu or weinberg@panther.adelphi.edu

Vol. 59, No. 7, 595–613 DOI: 10.1037/0003-066X.59.7.595

Drew Westen

that Meehl himself practiced psychoanalysis, despite his awareness of its inadequate evidentiary basis in replicable scientific studies (Meehl, 1978; Meehl, personal communication, 2002). Similar sentiments can be seen across the landscape in contemporary clinical psychology, as in the shift to clinical scientist models of clinical psychology training that minimize the importance of clinical experience for understanding clinical phenomena (e.g., McFall, 1991); models of treatment that minimize the role of clinical judgment on the grounds that such judgment is inherently inferior, over the long run, to interventions prescribed in a well-validated manual (see Westen, Novotny, & Thompson-Brenner, 2004); and models of assessment and diagnosis that advocate that clinicians replace their standard diagnostic practices with structured interviews that inquire about each diagnostic criterion for each disorder in the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM–IV*; American Psychiatric Association, 1994; Basco et al., 2000; Segal, Corcoran, & Coughlin, 2002; Wood, Garb, Lilienfeld, & Nezworski, 2002). Underlying all of these contemporary incarnations of the clinician–researcher tension that has existed since the rise of clinical psychology (see, e.g., McReynolds, 1987) is the view that clinical observations, judgments, procedures, methods of inquiry, and theoretical and technical predilections—to use Meehl's (1960, p. 19) term, the "cognitive activity of the clinician"—cannot be trusted.

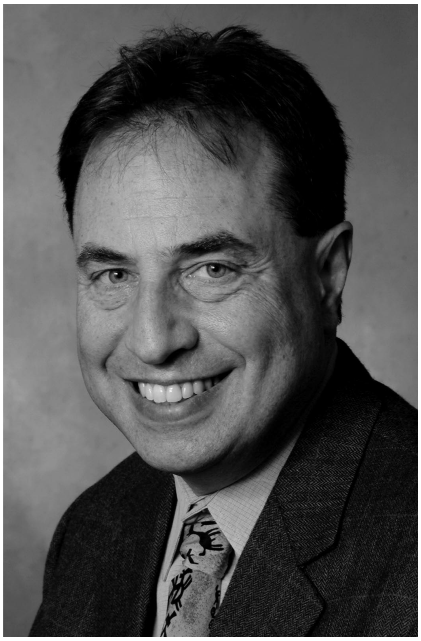
Our goal in this article is to revisit the clinical–statistical debate and, in the process, to rethink the question of what clinicians can and cannot do. We suggest that Meehl's arguments against informal aggregation stand 50 years later, but they have no bearing on whether, or under what circumstances, clinicians can make reliable and valid observations and inferences. We first address the dual meanings of the term *clinical* and examine the conditions under which the two types of clinical judgment are likely to be useful in prediction. We then review an emerging body of research on the quantification of clinical observation that considers what happens when we unconfound the two meanings, crossing clinical observation with statistical prediction. We conclude by reconsidering a paradox with which Meehl struggled throughout his career, a paradox that (in his words, cited above [Meehl, 1954, p. vi]) "besets most of us who do clinical work but try to be scientists," of how to reconcile idiographic (and potentially idiosyncratic) clinical judgment in a given hour with nomothetic science. We suggest that the clinical–statistical distinction constitutes as much a continuum as a dichotomy, and that every application of nomothetic, probabilistic statements to a given case (whether that case is a patient, a study to be designed or interpreted, or a body of literature) inherently involves clinical modes of aggregation.

Before proceeding, we should briefly note the potential meanings of the other word in the phrase clinical prediction, namely *prediction.* Cognitive processes can be arrayed on a continuum, from lower level processes, such as sensation and perception (which nevertheless involve substantial top-down processing), through processes denoted by terms such as inference, judgment, and decision making. Clinical observation includes substantial elements of perception and low-level categorization (e.g., the patient cries a lot or has a history of arrests) that require minimal inference. It also, however, includes substantial elements of judgment or inference (e.g., the patient is emotionally labile or is sensitive to rejection), which are not dissimilar in kind from the inferences required of lay observers when self-reporting symptoms or personality traits (e.g., "My mood is very changeable" or "I often worry about being rejected by people important to me"). As we argue below, there is good reason to believe that clinicians can make reliable judgments at this level of abstraction, which we denote here by the terms *observations*, *inferences*, and *judgments*. We restrict the term *prediction* to the way it is usually operationalized in research on clinical and statistical prediction, to refer to broader generalizations or prognostications (whether about past, concurrent, or future events), such as whether the patient is likely to have a history of sexual abuse or to make a successful suicide attempt in the next 2 years.

Two Meanings of *Clinical*

In an article on the "Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy" (Grove & Meehl, 1996), Grove and Meehl offered what may have represented Meehl's final word on the subject:

> Given a data set about an individual or a group (e.g., interviewer ratings, life history or demographic facts, test results, self-descriptions), there are two modes of data combination for a predictive or diagnostic purpose. The clinical method relies on human judgment that is based on informal contemplation and, sometimes,



Joel Weinberger
Photo by Mark Moss

> discussion with others (e.g., case conferences). The mechanical method involves a formal, algorithmic, objective procedure (e.g., equation) to reach the decision. Empirical comparisons . . . show that the mechanical method is almost invariably equal to or superior to the clinical method. (p. 293)

What is clear from this quote, and from his writings from 1954 onward, is that Meehl had in mind the distinction between two ways of aggregating data to make predictions or diagnoses, one highly inferential and synthetic, and the other mechanical or actuarial. At the same time, however, the term *clinical* connotes, if not denotes, a second distinction in psychology, between clinicians, who purport to have expertise in making judgments in a given domain, and nonclinicians, who claim no such expertise.

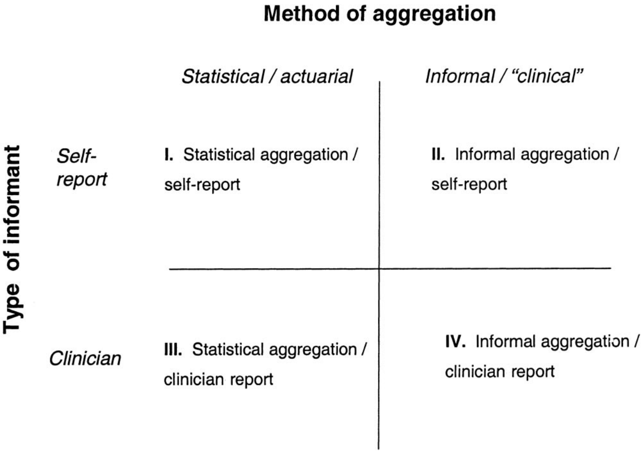
When Meehl first framed the debate, these two meanings of clinical were difficult to disentangle in psychology and psychiatry. Clinicians had tremendous latitude in making diagnoses. The first edition of the *Diagnostic and Statistical Manual* (*DSM*; American Psychiatric Association, 1952), the official set of rules for classifying mental disorders, provided few guidelines for aggregating clinical diagnostic data (see Spitzer, Endicott, & Robins, 1975). As a result, clinicians in different countries, cities, or even hospitals tended to use their own, often implicit diagnostic criteria. Psychoanalysis was also in a period of ascendance, enthusiasm, and hubris. It offered a mélange of theories, techniques, interpretive methods, and diagnostic distinctions, all derived exclusively from clinical observation. The lines were clear between practice and science, and by extension between clinical and statistical claims on knowledge.

Today, we may be in a better position to distinguish between these two meanings of clinical, as illustrated in Figure 1. The situations with which we are most familiar in clinical and personality psychology lie in the first and fourth quadrants. In the first quadrant, researchers aggregate psychometric self-report data statistically, as when they predict the likelihood of a future depressive episode from a patient's Beck Depression Inventory score (BDI; Beck, Steer, & Brown, 1996). In the fourth quadrant, clinicians aggregate clinical interview or other data informally, as when a clinician working with a patient assesses the probability that the patient will relapse if treatment is discontinued. These are the familiar quadrants of contemporary research and practice and the two poles most often associated with the clinical–statistical debate.

Until recently, virtually no research has addressed the third quadrant, which crosses clinical observation with statistical aggregation. (Quadrant II, in which undergraduates or psychiatric patients make unstructured judgments, has also received little empirical attention, precisely because of the recognition that such unstructured observation is unlikely to perform as well as standardized instruments in quantifying self-reports.) We hope to show that psychologists may have overestimated the deficiencies of clinical judgment by focusing on the reliability and validity of clinical judgment in Quadrant IV (broad, unstructured prognostications or dichotomous diagnostic judgments) rather than Quadrant III (statistical aggregation of clinical inference).

We are not, of course, the first to make distinctions between the method of aggregation and other variables relevant to the clinical–statistical debate. Meehl (1954) distinguished between the method of aggregation (clinical vs. statistical) and the type of data being aggregated (psychometric vs. nonpsychometric), irrespective of whether the data were provided by expert or lay observers. Dawes et al. (1989) similarly distinguished clinical methods of collecting versus interpreting information and explicitly focused only on the latter. Sawyer (1966) and later Wiggins

Figure 1
Method of Aggregation × Type of Informant



Type of informant	Method of aggregation: Statistical / actuarial	Method of aggregation: Informal / "clinical"
Self-report	**I.** Statistical aggregation / self-report	**II.** Informal aggregation / self-report
Clinician	**III.** Statistical aggregation / clinician report	**IV.** Informal aggregation / clinician report

(1973) reframed the debate in a way relevant to the current argument, distinguishing the method of aggregation and the method of measurement (which could be clinical, actuarial, or both). Sawyer foreshadowed the current argument when he suggested that "the clinician is more likely to contribute through observation than integration" (1966, p. 178). This suggestion, however, apparently did not resonate with researchers in this area, as evidenced by the virtual absence of subsequent data bearing on it (i.e., statistical aggregation of standardized clinical observations, using the kinds of psychometric methods developed for self-reports). What makes the distinction between the method of aggregation and the observer particularly important is that it falls along the fault line the debate tends naturally to take—and that has shown itself repeatedly throughout the history of our field—between clinicians, who claim to know something by virtue of their immersion in relatively unstructured clinical observation; and researchers, who view such claims as illusory in the absence of statistical data.

The Pitfalls of Clinical Aggregation: When Clinical Prediction (Frequently) Fails

Having identified two distinct meanings of clinical, we now examine each in turn. Meehl's original argument, that multiple regression is not easily done in one's head, is unassailable. As a result, statistical methods applied to most forms of data will produce results at least as good as subjective predictions. Under what conditions clinical prediction will fare better or worse depends in large measure on the answers to four questions: (a) How structured is the item set? (b) How did the judge (clinician or formula) combine the data to reach a judgment? (c) How many times has the judge confronted the task before, and to what extent has the judge received feedback and cross-replicated predictive algorithms? and (d) Does the task match the judge's experience? We address each of these questions only briefly, as they have received considerable attention in one form or another elsewhere (see Goldberg, 1991; Holt, 1958; Meehl, 1954; Westen & Weinberger, in press; Wiggins, 1973).

The first question pertains to the nature of the variables being aggregated. To make valid predictions, a judge (whether a person or an equation) needs equivalent data from one case to the next. Without it, the judge cannot develop weights—either informal or statistical—to apply to data in subsequent cases. If the variables included in an equation were different for each case, the equation could not generate valid predictions. We should not expect more of clinicians, and they should not advertise more.

The second question pertains to how the judge aggregates the data. Consider a study on the validity of personality disorder (PD) diagnosis using the LEAD standard (longitudinal evaluation using all available data; Spitzer, 1983). To make a LEAD diagnosis, multiple members of an investigatory team with knowledge of the patient from different sources and at different times (e.g., from structured interviews, observations on the ward, and informants) meet to arrive at a consensus diagnosis. They do this by evaluating each symptom of the diagnostic manual for each disorder and then applying the algorithms specified in the manual to make a diagnosis. In the absence of a gold standard for diagnosing PDs, this method, though flawed, is widely viewed as the next best thing. However, neither self-report questionnaires nor structured interviews for assessing PDs show substantial concordance with LEAD diagnoses (e.g., Perry, 1992; Pilkonis et al., 1995; Wilberg, Dammen, & Friis, 2000), raising the question of which, if any, of these methods best approaches diagnostic gold.

In one of the few studies comparing the predictive validity of LEAD and structured interview diagnoses, Pilkonis, Heape, Ruddy, and Serrao (1991) compared the outcome of depressed patients with and without a PD diagnosis according to LEAD consensus versus a well-validated structured interview. LEAD diagnosis predicted whether the patient was depressed six months later; structured interview diagnosis did not. Now imagine what would have happened if the researchers had asked the clinical team not to make consensus judgments on each of roughly 80 diagnostic criteria but to answer a single question: "How likely do you think this patient is to relapse within six months?" We suspect that actuarial prediction using interview data alone would have out-predicted LEAD diagnosis, because clinicians likely have no idea what algorithms reliably predict symptom change over time, any more than do patients responding to questions on the BDI or a structured PD interview. Indeed, the recent literature on affective forecasting indicates that people are vulnerable to enormous biases when asked to predict their future affective states (Gilbert & Ebert, 2002). What distinguishes clinical prediction in this study from the prototypical study of clinical versus statistical prediction is, first, that clinicians reached consensus not on a single statement, but on approximately 80; and second, that multiple clinicians came to the case conference with independent assessments of each item. Reliability theory would suggest that as the number of both items and raters increases, so should reliability of measurement (see Cronbach, Rajaratnam, & Gleser, 1964; Epstein, 1986; Meehl, 1960; Strube, 2000).

A third question pertains to the number of times the judge has confronted the task before and the extent to which the judge (clinician or formula) has had the benefit of feedback and cross-validation using different samples (see Holt, 1958). As noted by numerous commentators since Meehl (e.g., Dawes et al., 1989), many clinical decisions are made repeatedly without possibility of self-correction because clinicians never receive feedback about outcomes. Imagine the analogous situation for statistical prediction, if a researcher designing a test to predict mania never saw whether patients who completed the procedure became manic. Without testing initial items and weights against known criteria and refining those items and weights over successive iterations with different samples to minimize sources of variance idiosyncratic to one or another data set, researchers could not build predictive equations.

A final question is the extent to which judges are making inferences about questions (and samples) for which they have expertise (Holt, 1958). One would expect that the impact of clinical training would be most apparent when clinicians are answering the kinds of clinical questions they are called upon to answer on a daily basis, such as whether a person appears to have hallucinations or is prone to self-criticism. We would not expect clinicians to be expert in making judgments about whether a person is likely to succeed in the Peace Corps (see Mischel, 1968).

Thus, a clinician whose goal is valid prognostication would do well to rely on a standard set of items, make judgments at an appropriate level of inference that capitalizes on skills likely to have been developed through clinical training and experience, make multiple such judgments that can then be aggregated, and avoid prognosticating outside of his or her area of expertise, except where statistical prediction would be premature because of lack of information or inadequate knowledge of relevant variables and their relative contributions. This last point, however, raises an important caveat. We may undervalue the utility of clinical observation in prediction if we assume an item set with useful predictors and compare it to a clinician. As Meehl believed (personal communication to Drew Westen, September 2000), there is no substitute for clinical experience in generating hypotheses and devising clinically relevant items for use in research. Consider the concept of psychopathy, a precursor to the *DSM–IV* antisocial PD diagnosis. The psychopathy construct is currently experiencing a renaissance (and a likely return in some form to a future *DSM*) because it tends to be more predictive of outcomes than the antisocial diagnosis, which focuses more on antisocial behaviors and less on underlying personality dispositions (e.g., Hare, 1998; Lorenz & Newman, 2002). Virtually all current research on psychopathy, however, presupposes the observations of a brilliant clinical observer (Cleckley, 1941), whose clinical immersion among psychopaths over 60 years ago still provides the foundation for the measure considered the gold standard in psychopathy research (Hare et al., 1990). Had Cleckley not identified and aggregated a set of important variables in the best sense of "clinical" intended by Meehl, we would *have* no statistical prediction.

What Meehl and others (e.g., Dawes et al., 1989) have appropriately argued is that clinicians too frequently ignore or override statistical data, assuming either that they have some special skill that allows them to outperform formulas or that "every case is different," a war cry that would invalidate all prior knowledge, clinical or statistical. Furthermore, clinicians are prone to the same heuristics and biases that plague lay judgments, inferences, and prognostications (e.g., Goldberg, 1991; Kahneman & Tversky, 1973, 2000; Nisbett & Ross, 1980) and remain vulnerable to these biases unless they are aware of them and exercise appropriate vigilance. On the other hand, there is often no algorithm or objective method for determining when to apply a formula to a particular case. Indeed, as addressed below, the decision to *use* a formula in a given instance is as "clinical" in Meehl's sense as the decision to reject it.

The Nature of the Observer: Clinician Reports Versus Self-Reports

We now turn to the second meaning of clinical and address the question of whether or under what conditions one does better to rely on the observations of expert clinical observers or non-expert observers. Research in cognitive science suggests that with increasing experience in a given domain, people are typically able to make more subtle discriminations, process information more efficiently, and automatize procedures that initially required conscious attention and hence consumed working memory resources. With expertise, basic-level concepts (e.g., chair) become too basic for thought and discourse, and concepts considered subordinate in lay categorization (e.g., Queen Anne chair) tend to function like basic-level concepts (e.g., Tanaka & Taylor, 1991). Thus, we would be surprised (and concerned) if our automobile mechanic shared our lay diagnosis that "there's a clanking sound under the hood."

Yet as a field, we rely heavily on lay informants. The self-reports of undergraduates and psychiatric patients constitute the vast majority of data in personality and clinical psychology, whether assessed directly by questionnaire or more indirectly by structured interview. Unfortunately, we know very little about how expert versus lay observations of personality or psychopathology fare in predicting a range of outcomes using a variety of methods of aggregating those observations (including actuarial methods, as in Quadrant III of Figure 1). Here, however, we briefly summarize some of the advantages and disadvantages of the two kinds of observers, beginning with self-reports, and consider the kinds of situations in which we might expect one, the other, or both to be useful.

Advantages and disadvantages of self-reports. The advantages of self-reports are well known. First, for many questions, people are the most obvious source of data about themselves because they have the widest observational base. If we want to know how much someone thinks about suicide or enjoys interacting with people, we do well to start at the source. Second and related, if we want to know people's explicit beliefs or memories for a particular event or set of events (their conscious phenomenology), we should ask them. Third, from a pragmatic view, self-reports are easy to obtain, and to the extent that they account for a substantial percentage of variance in assessing a given construct, their benefit-to-cost ratio will be high. Fourth, empirically, self-reports have paid off. The advances in the behavioral genetics of personality over the last three decades are a testament to the value of well-constructed, well-validated self-report instruments (e.g., Harkness, Tellegen, & Waller, 1995).

Self-reports also, however, have limitations. We note four (Block, 1995; McAdams, 1992; Westen, 1995, 1996):

1. Understanding personality and psychopathology presumably requires training and experience (or so licensing boards believe, rightly or wrongly), just as does understanding of automobiles or infectious diseases. We suspect most readers would be taken aback if they brought their car in for repair, and instead of opening the hood, the mechanic

asked them to complete a problem checklist in lay language ("It won't start," "There's black smoke coming out of it," "The heater won't turn on"), calculated factor scores (low on clanking, low on starting, high on tire pressure), and proceeded to install a new transmission. Nor would most readers likely be reassured if the mechanic cited evidence that the factor scores show high test–retest and interrater reliability (the same person typically reports the same problem two days in a row, and both members of a couple who drive the car tend to concur on the problem) and correlate .30 with relevant criterion variables (including improvement with a new transmission). The mind is surely as complex as an automobile engine, and it is difficult to imagine that lay observation and item content designed to minimize intellectual and literacy requirements (e.g., items written at a sixth-grade reading level) are always sufficient for making observations necessary for subtle diagnostic and predictive judgments.

2. As Nisbett and Wilson (1977) showed over a quarter century ago, people have minimal access to many of their cognitive processes, and they often confabulate explanations for their behavior by applying intuitive attributional theories ("I guess I did that because . . ."). Research since that time has demonstrated that much of human behavior reflects consciously unreportable (implicit) rather than reportable (explicit) processes, and that this applies to virtually every area of psychological functioning, including memory, cognition, emotion, attitudes, and motivation (Weinberger, in press; Westen, 1998; T. D. Wilson, Lindsey, & Schooler, 2000). For example, McClelland, Koestner, and Weinberger (1989) showed that explicit (self-report) and projective (implicit) measures of motives do not correlate with each other, but that each has theoretically and ecologically meaningful correlates. Implicit motives express themselves across long periods of time and can be activated without conscious awareness, whereas explicit motives influence behavior only when conscious attention is drawn to them. Psychopathology researchers have similarly begun to exploit the distinction between implicit and explicit processes using procedures such as emotional Stroop tasks that access implicit attentional biases (e.g., Williams, Mathews, & MacLeod, 1996). To the extent that personality or psychopathology variables are not accessible to introspective awareness, they will not be accessible by self-report.

3. Self-reports can be limited by defensive and self-presentational biases, which social–psychological research on self-serving biases suggests are extensive (Epstein, 1992; Greenwald, Pratkanis, Leippe, & Baumgardner, 1986; John & Robins, 1994; Paulhus, Fridhandler, & Hayes, 1997). The majority of people describe themselves as above average on the majority of traits the majority of times. This is in fact one of the few areas in which researchers have demonstrated incremental validity of ratings made by trained observers ("clinicians," or purported experts, in Meehl's sense) relative to self-reports. For example, in two studies, Shedler, Mayman, and Manis (1993) examined participants who reported themselves to be free of psychological distress and symptomatology but whose narrative descriptions of their early memories were rated by an experienced clinician as showing signs of psychological disturbance. While undergoing a mildly stressful procedure, participants who viewed themselves as healthy but who the clinician had identified as distressed showed significantly greater cardiac reactivity than patients who were either low or high on both measures of distress. They also showed more indirect signs of anxiety (such as stammering, sighing, and avoiding the content of the stimulus) while reporting less anxiety than other participants. Of particular import, self-report scales designed to detect self-presentational biases were unable to detect these individuals (Shedler, Mayman, & Manis, 1994). In another study, students who showed this pattern of low self-reported and high clinician-reported negative affect made more health care visits over the next year than those who admitted their distress (Cousineau, 1997). Interestingly, the results were much weaker for self-reported visits to the doctor than for documented visits, suggesting the extent to which defensive biases can affect even seemingly objective criteria (and produce spurious conclusions about the benefits of such distortions when the predictor and criterion variables share common error variance; see Colvin, Block, & Funder, 1995).

4. In most areas of psychology, we measure skills or aptitudes rather than asking individuals to self-report them. We do not measure intelligence by asking participants to make 5-point Likert-type ratings of items such as, "I know a lot of big words" or "I can picture things better than most people." Rather, we observe their performance on relevant tasks. We suspect the correlation between self-reported and observed vocabulary equals or exceeds the typical personality coefficient of .30, but most of us would not substitute IQ by self-report for IQ functionally assessed through behavioral observation. Although one could make arguments for the likely superiority of self-reported personality over self-reported intelligence, we suspect that the major reason we assess individual differences so differently in the two domains largely reflects factors specific to the history of the two subdisciplines.

Advantages and disadvantages of clinician reports. Data provided by clinician informants have advantages and disadvantages as well. With respect to advantages, first, clinicians are experienced observers, whose observations and inferences reflect years of training and experience. By virtue of their experience, they are also likely to have a normative basis from which to make inferences about psychopathology. Their implicit norms may differ from one another and hence reduce reliability, just as patients' implicit norms influence their responses. Nevertheless, we would expect individuals who have seen dozens of depressed or psychotic patients to be able to make finer and more reliable discriminations than lay observers (particularly when these lay observers are mired in their own depression or psychosis). Instruments devised for expert report also need not be written at a reading level (typically sixth grade) that constrains the constructs that can be assessed by self-report.

Second, to the extent that clinicians observe important aspects of patients' behavior directly, their observations are likely to have value added relative to self-reports. In the assessment of personality pathology, clinicians of every theoretical orientation gravitate toward two assessment methods: observing patients' behavior in the consulting room and listening to their narratives of significant events (particularly interpersonal events; Westen, 1997). Clinical consensus is by no means an index of the validity of an assessment procedure, but when clinicians with highly disparate professional training (psychologists vs. psychiatrists) and theoretical orientations gravitate consistently in one direction, we should at least consider the possibility that they are on to something other than shared error. In fact, an emerging body of evidence suggests that direct observation of interpersonal behavior and attention to structural qualities of narratives provide not only useful but incremental information above and beyond individuals' explicit self-reports in a number of domains, such as adult attachment (Dozier & Kobak, 1992; Fonagy, Steele, & Steele, 1991; Main, Kaplan, & Cassidy, 1985).

The third advantage of clinicians as informants is that they do not share patients' defensive and self-presentational biases—biases that can be particularly problematic when patients are asked to describe socially undesirable or embarrassing symptoms or traits (Thomas, Turkheimer, & Oltmanns, 2003). Clinicians, of course, have their own biases (an issue we address below). However, at the very least, this *different* source of error should make clinician reports a useful complement to self-reports. Presumably patients who lack insight do not uniformly work with clinicians who lack insight.

Finally, for research in psychopathology, clinician reports have the same advantage as self-reports: ready accessibility. The accessibility of clinicians, each of whom is likely to see 20 to 40 patients per week, makes possible large sample sizes that can be essential for many forms of psychiatric research, such as taxonomic research. Researchers can collect a sample of 500 or 1,000 patients from a random national sample of PhDs and MDs in a matter of months by accessing databases of clinicians from the registers of appropriate professional organizations. We describe some examples of this approach below.

Although instruments designed to quantify clinical observations might provide a useful complement to more traditional questionnaire and interview methods, they have limitations as well. We focus here on both these limitations and on their boundaries, given widespread skepticism about clinicians as potential informants, which we believe largely reflects a misunderstanding of the clinical–statistical literature.[1]

A first concern is that clinicians' theoretical biases could influence their observations. Clinicians do have biases, as do all observers. The extent to which these biases are larger or more systematic than the biases imposed by lay informants' intuitive psychological theories is unknown. Fortunately, the impact of such biases can be tested by obtaining theoretically and professionally diverse samples (e.g., psychiatrists and psychologists; clinicians with a range of theoretical orientations). Empirically, we have found surprisingly little evidence of theory-driven observational bias in research using clinician-report methods for a range of disorders, even in assessing highly theory-driven domains. For example, as part of a broader study validating a set of clinician-report measures, Betan, Heim, Zittel, and Westen (2004) recently administered a 79-item clinician-report measure of "countertransference processes" broadly construed (referring to feelings elicited in the course of working with a given patient). Participants comprised a random national sample of doctoral-level clinicians describing a randomly selected patient in their care. Factor analysis using the entire sample yielded eight factors; a second factor analysis deleting all clinicians who self-reported a psychodynamic orientation (from which the concept of countertransference emerged) yielded precisely the same factor structure. In other research, asking clinicians to describe the personality of a patient they have diagnosed with a particular PD does not yield descriptions that closely mirror *DSM–IV* criteria even when clinicians are aware of those criteria, suggesting that clinicians, when asked to describe a specific patient using specific items, tend to describe what they have observed rather than to recount a diagnostic prototype (Shedler & Westen, 1998, 2004; Westen & Shedler, 1999a).

A second concern is that clinical observations, because they are not based on structured interviews, reflect unknown and variable data acquisition strategies. Like the limitation of theory-driven biases, this is a genuine limitation but is, once again, important to consider in the context of the limitations of more traditional methods. Clinicians listen to patients' narratives of emotionally charged events contemporaneously over time (typically within days of the events), observe their symptoms wax and wane over time, observe their behavior in what often becomes an emotionally important relationship, and so forth. In the case of children and adolescents, clinicians also frequently have numerous interactions with parents, schools, and collateral and past caregivers. In contrast, participants in psychiatric research typically see an interviewer (if at all) on only one occasion, have unknown or variable motivation to self-disclose, and provide responses in a very structured setting with substantial time constraints. The modal interviewer in psychiatric research is a bachelor's-level research assistant with little clinical exposure beyond initial training using a particular structured interview (see, e.g., Kranzler, Kadden, Babor, Tennen, & Rounsaville, 1996), who may or may not

[1] For readers who are unconvinced that such biases and misunderstandings are widespread, consider the following critique of a manuscript that relied on psychometric data provided by clinicians rather than patients: "It is difficult to ignore a large body of evidence, dating back over 30 years, regarding biases in the types of judgments that the clinicians in this study made." The author of this statement is editor of one of the major American Psychological Association (APA) clinical psychology journals. In our experience, this response is not unusual (at least among scientists who review for APA journals). It is rapidly becoming modal among academic clinical psychologists, who, unlike Meehl, tend to believe that they (and others) have little to learn from clinical practice or experience and whose attitudes tend to reflect what might be called *clinicism* (cynicism toward, and negative stereotypes of, clinicians).

recognize subtle verbal, postural, or behavioral indicators suggesting the need for further probing, questioning the fidelity of the patient's report, and so forth (cf. Brammer, 2002).

With respect to reliability of such judgments, the limitations of clinician reports are no greater than those of self-report questionnaires, which require the same kind of unstructured data aggregation or generalization at the item level as clinician reports. When a patient responds to an item assessing her tendency to feel depressed, she must think back across instances or consult her prototypic self-concept, just as the clinician must abstract over episodes or consult a mental prototype of the patient. A considerable body of evidence suggests that even minor changes in wording or context can have an enormous impact on the way people aggregate data at the item level (Schwartz, 1999). Structured interviews also rely heavily on self-reports with unknown reliability and validity, and it is precisely because these interviews allow some measure of clinical inference that most researchers consider them preferable to questionnaires for establishing diagnoses in psychiatric research.

The use of clinician reports, like self-reports, does not require the reliability of any given data point. A felicitous consequence of the large sample sizes made possible by using self-reports with undergraduates—and clinician reports in studies of psychopathology—is that, even with measures with only moderate reliability (or low reliability at the item level), randomly distributed errors around a mean are likely to provide measures of central tendency that are as or more reliable than those obtained using more reliable assessment methods applied to small samples (see Rosnow & Rosenthal, 1991). As generalizability theory suggests (Cronbach et al., 1964; Strube, 2000), one can maximize reliability and generalizability in multiple ways, ranging from intensive interviewer training to increasing the number of items, raters, or participants whose data are pooled, as long as errors are uncorrelated (or can be controlled statistically, e.g., if researchers find a relation between shared training or theoretical orientation and responses).

Summary: The nature of the informant in the clinical–statistical debate. The question of when to rely on self-reports versus clinician reports seems to us at once theoretical, empirical, and pragmatic. Theoretically, self-reports are likely to provide valid data when participants are describing behaviors or mental processes that require minimal expertise, are readily observable to themselves or readily generalized (e.g., how many friends they regularly see, how often they spend time with their family, whether they get upset being alone, whether they enjoy the opera, whether they have made suicidal gestures), and have minimal bearing on self-evaluation (and hence are less likely to activate defensive and self-presentational biases). Thus, people are usually able to report accurately on the extent to which they are extroverted—unless they are manipulated to believe that extroversion is bad (Kunda, 1990). When these conditions are violated, researchers should turn to data from other informants.

In contrast, clinician reports are likely to be most useful when responses require experience with psychopathology. They are also useful when the domain being assessed, even if inaccessible to self-report, has manifestations in behavior that can be decoded (implicitly or explicitly) by an experienced observer; when the population of interest is represented among patients seen in clinical practice settings or can be represented using targeted sampling strategies (e.g., collecting data from clinicians who work in forensic settings to study psychopathy); and when the clinician knows the patient relatively well. Clinician reports also have the same pragmatic advantage in studying psychopathology that self-reports have in studying personality, namely the ready accessibility of informants. As with self-reports, when these conditions are violated, researchers should seek alternative informants.

In light of the arguments above, the absence of compelling data on the advantages conferred by clinical training and experience beyond, perhaps, a year or two of graduate school (see Garb, 1998) is surprising, and particularly so for three reasons. First, relative to self-reports, the use of clinical informants has the same advantage as the use of lay (non-self) informants: Observations can be aggregated across multiple observers to maximize reliability (see, e.g., Block, 1971; Block & Block, 1981). Oltmanns, Turkheimer, and their colleagues (e.g., Fiedler, Oltmanns, & Turkheimer, in press; Thomas et al., 2003) have shown that aggregated peer descriptions of PD traits can be extremely reliable and show incremental validity in predicting real-world outcomes (e.g., whether a military recruit completes his or her intended term of military service), holding constant self-reports of the same constructs. Peer reports of personality pathology actually appear to be substantially better predictors of completion of military service than self-reports, in large measure because of their stronger validity in assessing externalizing pathology (Fiedler et al., in press). Ready, Watson, and Clark (2002) found that self- and observer reports each predict unique variance in criterion variables related to personality pathology. That aggregated clinician reports could not capitalize on the same psychometric principles as aggregated data from college students or cadets seems unlikely.

Second, data coding strategies that mirror what clinicians do in making judgments about personality—observe patients' behavior and listen to their narratives for elements of syntax, prosody, content, and so forth that might yield information about who they are—often show substantial reliability and validity, including incremental validity vis-à-vis self-reports. For example, attachment researchers have demonstrated that data provided by trained raters coding narratives can predict an unborn infant's attachment security at 12 to 18 months (Fonagy et al., 1991), and that narrative-based and self-report attachment measures are largely uncorrelated but each predicts substantial variance in attachment-related criterion variables (Cassidy & Shaver, 1999). Dozens of studies similarly support the predictive validity of reliably coded open-ended verbal or narrative responses. These include Loevinger's Sentence Completion Test (Loevinger & Wessler, 1970), Rorschach data assessed using Holzman's Thought Disorder Inventory (Johnston & Holzman, 1979), Thematic Apperception Test measures of implicit motives (McClelland, 1985; Smith, Atkinson, McClelland, & Veroff, 1992), and a host of other narrative-based measures that require expert (in Meehl's sense, clinical) inference (see Westen, Feit, & Zittel, 1999).

Third, clinicians would be a very peculiar species indeed if they showed no skill development over years of observing and treating psychopathology.[2] Research on implicit learning suggests that people learn all kinds of regularities about their environment that become expressed in skilled performance even when they have no explicit knowledge of the implicit "grammars" generating their inferences or actions (e.g., Reber, 1992; Rubin, Wallace, & Houston, 1993). Lewicki (1986) showed that both adults and children can unconsciously learn quite complex covariations among social stimuli that they are unable to report explicitly. It would be remarkable if clinicians could not also do so. Lewicki was able to measure such implicit learning with lay adults and children. Psychologists should be able to devise ways to accomplish this with clinicians. We now describe a research program that has attempted to do so.

Clinical Observation × Statistical Aggregation: The Missing Interaction Term?

It is also possible that interview-based judgments at a minimally inferential level, if recorded in standard form (for example, Q-sort) and treated statistically, can be made more powerful than such data treated impressionistically as is currently the practice. (Meehl, 1959, p. 124)

In this section, we examine what quantified judgments made by clinical informants might be able to do when aggregated statistically (Quadrant III of Figure 1). Over the last several years, Westen, Shedler, and colleagues have developed a set of measures to quantify the observations of clinician informants, using the same psychometric procedures personality and clinical psychologists have used over the last five decades to quantify the observations of lay observers. In line with Meehl's suggestion above, Shedler and Westen developed an omnibus Q sort instrument called the SWAP–200, akin to a clinician-report Minnesota Multiphasic Personality Inventory–2 (MMPI–2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), to assess personality pathology in adults (Shedler & Westen, 2004, in press; Westen & Shedler, 1999a, 1999b) and adolescents (Westen, Shedler, Durrett, Glass, & Martens, 2003). Since that time, Westen and colleagues (we use the term "we" here for simplicity) have developed instruments assessing a range of more specific personality variables in greater detail (essentially magnifying dimensions covered by the SWAP–200 to allow more fine-grained distinctions), such as emotion regulation and emotional experience (Westen, Muderrisoglu, Fowler, Shedler, & Koren, 1997), identity disturbance (Wilkinson-Ryan & Westen, 2000), impulsivity (e.g., Novotny, Eddy, & Westen, 2004), and subclinical cognitive disturbances (Heim & Westen, 2002). More recently, we have developed and begun piloting clinician-report measures of eating, mood, anxiety, and substance use disorders.[3] Aside from the potential usefulness of these measures in basic science (notably taxonomic) research, our goal is to support a more scientific attitude to clinical practice, by allowing clinicians to diagnose and track change in key areas over time, using normed instruments instead of intuitive judgments (see also Stricker & Trierweiler, 1995). Such instruments could also prove useful in psychotherapy research using large practice networks.

In this section, we first address the question of whether clinicians can provide reliable and valid data if asked to make observations and inferences at a level of generality that maximizes the usefulness of clinical expertise. We then present an example of research aimed at refining the classification of psychopathology by statistically aggregating clinician-report data.

Structure and Correlates of Clinician- and Lay-Report Data

One way of assessing the ability of clinicians to provide psychometrically reliable and valid data is to examine the factor structure, reliability, and external correlates of their ratings using measures with well-known psychometric properties. This allows us to gauge the extent to which clinician reports show similar operating characteristics to quantified judgments made by other informants. For illustration, we describe a recent study (Dutra, Campbell, & Westen, 2004) in which a large sample of clinicians each described the psychopathology of an adolescent patient in their care using the parent version of the Child Behavior Checklist (CBCL; Achenbach, 1991).

The CBCL is a widely used questionnaire, first designed for parent report and then for teacher and self-report. It assesses the behavioral problems and social competencies of children 4 to 18 years of age. The CBCL is composed of 118 problem items and 20 competence items grouped into 11 Problem scales (including two broadband factors, Internalizing and Externalizing*)* and 4 Competence scales. (We focus here only on the Problem scales, because they are most relevant to psychopathology and were the focus of our study.) The CBCL is broadly used in both clinical and research settings because of its demonstrated reliability and validity and broad applicability across ages and populations.

[2] To what extent clinicians obtain feedback relevant to developing expertise is, of course, an open question. Although psychologists who limit themselves to diagnostic testing without longitudinal follow-up may well calcify their biases over time, and all clinicians tend to elicit material from patients that fits their preferred theories (much as researchers elicit data from statistical procedures that fits their predilections), patients are hardly so pliable and suggestible that clinicians receive no ongoing feedback. In our clinical experience, patients routinely say things like, "No, I think you're misunderstanding me," "That's not really right," or "I don't think what we're doing is helping me." Indeed, psychotherapists tend to have much more direct and immediate feedback than most other medical practitioners, who may prescribe a medication or perform a procedure and not see the patient again for a year.

[3] Measures and published papers are available at www.psychsystems.net/lab

As part of a broader project on personality pathology in adolescents (described below), 294 clinicians, randomly selected from the registers of the American Academy of Child and Adolescent Psychiatry and the APA, completed the parent-report version of the CBCL on a randomly selected adolescent patient in their care. To see whether we could recover the factor structure of the instrument using clinicians as informants, we conducted a confirmatory factor analysis (CFA), replicating procedures (including item parceling, with two parcels for each of the eight lower order scales) used by Greenbaum and Dedrick (1998) in a CFA of the CBCL based on parent reports.

With respect to factor structure, the data strongly resembled the parent-report data: Each set of items loaded on the expected lower order factor (with virtually all factor loadings $> .70$), and the lower order factors in turn loaded as predicted on the higher order Internalizing and Externalizing factors. Internal consistency (reliability) for both the problem scales and broadband Internalizing and Externalizing factors was adequate in all but one case (median alpha $= .76$, range $= .55–.94$). This pattern of coefficients was similar to those obtained in studies using the CBCL with other informants (e.g., the Sex Problems scale tends to show lower alphas).

Validity data strongly supported a range of theory-driven hypotheses. For example, clinician ratings of school functioning were negatively associated with the Externalizing, Attention Problems, and Delinquent scales (all *r*s around –.50), and clinician ratings of quality and number of peer relationships were negatively correlated with the Social Problems and Withdrawn scales (*r*s ranging from –.30 to –.61). Measures of adaptive functioning requiring minimal clinical inference showed the same pattern of associations as more inferential ratings (e.g., arrest history was significantly correlated with the Delinquent Behavior and Aggressive scales). CBCL scores also showed predicted patterns of familial aggregation. For example, Thought Problems scale scores were specifically associated with a family history of psychosis, and Delinquent Behavior, Aggressive Behavior, and Externalizing scale scores were all specifically associated with a family history of alcoholism, illicit substance use disorders, and criminality.

The central implication of these findings is that, when using a well-understood, well-validated instrument with known psychometric properties, experienced clinicians do not show the kinds of biases and errors often attributed to them. Instead, they provide data with reliability and validity comparable to those of other informants, and their data show similar factor structure. We cannot tell from these data whether clinician reports are superior or inferior to parent, teacher, or youth self-reports using the CBCL in predicting a range of external criteria; this question awaits future research. What we do know, however, is that clinician-report adaptations of other well-validated measures are producing similar results (e.g., Russ, Heim, & Westen, 2003).

Convergence of Data From Treating Clinicians, Independent Interviews, and Self-Reports

In the CBCL study, some of the criterion variables were relatively objective and hence less vulnerable to bias (e.g., whether the patient had been hospitalized, made a suicide attempt, or had been arrested). However, like the vast majority of studies in personality and clinical psychology (which predict self-reports from self-reports), we could not conclusively distinguish rater variance from true variance because a single observer (the clinician) provided all the data. Two recently completed studies do not share this limitation and suggest that clinical judgment can be highly reliable and valid if quantified using suitable psychometric instruments.

Both studies used the SWAP–200 Q sort, which was designed for expert clinical observers. A Q sort (in the context of personality assessment) is a set of personality-descriptive statements that may describe a given person well, somewhat, or not at all. The statements are printed on separate index cards, and an observer with a thorough knowledge of the subject sorts (rank-orders) the statements into categories, from those that are inapplicable or not descriptive to those that are highly descriptive (see Block, 1978). The task of the observer using the SWAP–200 is to sort 200 statements into eight rank-ordered categories, from 0 (items judged irrelevant or inapplicable to the patient) to 7 (items deemed highly descriptive). The item set was developed, revised, and honed using standard procedures for item refinement used by personality psychologists, such as soliciting feedback from hundreds of clinicians who used the item set to describe their patients, eliminating items with minimal variance or high redundancy with others, and so forth.

Westen and Muderrisoglu (2003) interviewed a small sample of outpatients ($N = 24$) using the Clinical Diagnostic Interview (CDI; Westen, 2002). In contrast to a structured interview, the CDI is what might be called a systematic clinical interview, designed to mirror but systematize the kind of interviewing approach used by experienced clinicians of all theoretical orientations to assess personality (Westen, 1997). Although the CDI includes a number of direct questions (e.g., about characteristic moods), it does not primarily ask patients to describe their personalities. Rather, it asks them to tell narratives about their lives and problems that allow the interviewer to make judgments about their characteristic ways of thinking, feeling, regulating emotions and impulses, experiencing themselves and others, and so forth. The interview begins, as in a standard clinical interview, by asking patients what brought them to treatment, with the interviewer probing for details about severity, frequency, duration, and history of symptoms. The interviewer then asks patients about a wide range of significant relationships and experiences from the past and present (e.g., parents, siblings, romantic relationships, friendships, school and work experiences, particularly stressful times). For each category of relationship or experience, the interviewer asks the patient to describe two to three specific incidents. The interview assumes a competent, experienced clinical interviewer: Although the interview specifies a set of core questions that provides the skeleton of the interview (and suggests specific probes throughout), most probing depends on clinical judgment.

The primary aims of the study were twofold: to assess the reliability with which two clinician–judges, independently viewing the same interview, made SWAP–200 assessments using the CDI; and to assess the convergence between these interview-based assessments (aggregated across the two judges, to maximize reliability) and treating clinicians' SWAP–200 descriptions of the patient based on their contact with the patient over time. (All clinical judges, including the two CDI interviewers as well as the treating clinician, were blind to data provided by the others.) The variables of interest were patients' PD scale scores. To calculate SWAP–200 PD scores, patients' 200-item profiles are correlated with empirically derived 200-item prototypes of each diagnosis under consideration. For ease of interpretation, these correlations (between the patient's profile and each of several diagnostic prototypes) are then converted to T scores. (Patients can also receive traditional factor-based scores derived from conventional factor analysis [Shedler & Westen, in press], which showed similar results in this study.) The investigators assessed the reliability and validity of SWAP–200 diagnosis using both the *DSM–IV* PDs (correlating patients' profiles with aggregated prototypes derived from a national sample of patients with PDs) and a set of seven PD diagnoses empirically derived using a clustering procedure (Q factor analysis) in a prior large *N* sample.

Primary findings were as follows. Interrater reliability (two interviewers per patient) averaged greater than .80 for all Axis II and empirically derived diagnoses. This is noteworthy in two respects. First, the task of the judge is to sort 200 items based on the degree to which they are descriptive of the patient, following an extensive narrative-based interview that requires, rather than eliminates, clinical judgment. Second, diagnosis is strictly actuarial, reflecting the degree of match between the patient's 200-item profile and empirical prototypes. This method requires clinicians to make sophisticated clinical inferences, but it does not require them to aggregate those inferences to make diagnoses (particularly categorical diagnoses).

With respect to validity, median correlations between PD scores derived from the treating clinician's Q sort description of a patient and the interviewers' description of the same patient were greater than .80 for the 10 *DSM–IV* Axis II disorders as well as for the seven empirically derived diagnoses. Discriminant validity (correlations off the diagonal, between treating clinicians' PD scores for one disorder and interviewers' PD scores for another) was only moderate for *DSM–IV* diagnoses (median $r = .40$), which is unsurprising given the diagnostic redundancy built into the Axis II criterion sets. In contrast, the median correlation off the diagonal for the empirically derived PD diagnoses and factors hovered around zero.

Thus, in this preliminary study, we obtained highly reliable diagnoses among two interviewers; strong evidence of convergent validity, with diagnostic judgments made by the patient's clinician correlating highly with interviewer judgments; and strong evidence of discriminant validity when criterion diagnoses were empirically derived to minimize redundancy. As noted earlier, and by way of comparison, the correlations between structured interview diagnoses and LEAD diagnoses tend to range from .00 to .40, with poor discriminant validity (see Pilkonis et al., 1991, 1995). Similarly, a meta-analysis of the magnitude of self-informant correlations for PD dimensions assessed by structured interview and self-report questionnaires yielded a median correlation of .36 (which did not differ for interviews vs. questionnaires; Klonsky, Oltmanns, & Turkheimer, 2002).

A second study examined the relation between SWAP–200 descriptions made by the treating clinician and patient self-reports (Bradley, Hilsenroth, & Westen, 2003). Advanced graduate students in clinical psychology participating in a psychotherapy process-outcome study used the SWAP–200 to describe 54 outpatients after the fifth contact hour (including two hours of intake interviews loosely based on the CDI). Patients completed the Personality Assessment Inventory (PAI; Morey, 1991) and the Inventory of Interpersonal Problems (IIP; Horowitz, Rosenberg, Baer, Ureno, & Billasenor, 1988). Inclusion of these self-report measures allowed us to examine the convergence between clinician diagnoses using the SWAP–200 and self-reported borderline and antisocial features on the PAI (the two PDs for which self-informant convergence has tended to support validity of self-reports) and interpersonal problems assessed using the IIP.

The data provided further support for the validity of clinical inference using the SWAP–200 Q sort. For example, antisocial PD scores on the SWAP–200 differentially predicted antisocial and aggression scores on the PAI, whereas borderline PD scores on the SWAP–200 predicted borderline PAI scores. Quantified clinical judgment predicted scores on the IIP as well. For example, SWAP–200 antisocial scores predicted IIP Dominance, and schizoid and schizotypal PD predicted IIP Coldness and Avoidance. Taken together, these two studies suggest that clinicians can make reliable and valid diagnostic discriminations, based on either their clinical observation of a patient over the course of treatment or administration of a systematic clinical interview, if provided with a suitable psychometric instrument rather than asked to aggregate their inferences into dichotomous, present–absent diagnostic judgments.

A study in progress from a research group other than our own has just reported comparable findings (Marin-Avellan, McGauley, Campbell, & Fonagy, 2004). The investigators applied the SWAP–200 to audiotaped Adult Attachment Interviews (Main et al., 1985) plus chart records on a sample of inpatients at a maximum security forensic hospital (a method similar to methods for coding psychopathy; Hare et al., 1990). Thus far, the investigators have analyzed data from the first 30 cases of an ongoing study. Interrater reliability for SWAP–200 PD scale scores was high, with a median correlation of $r = .91$. However, the most important findings pertain to prediction of actual ward behavior. SWAP–200 PD scores proved superior to diagnosis using the Structured Clinical Interview for *DSM–IV* Axis II (SCID–II; First, Spitzer, Gibbon, & Williams, 1997) in predicting a range of interpersonal variables rated by nurses on the ward using a 49-item interpersonal circumplex rating scale. For example, antisocial PD as assessed by both instruments predicted dominance behavior on the ward; however, only SWAP antisocial scores predicted coercive behaviors. The SWAP, unlike the SCID–II, also yielded negative correlations between antisocial PD and submissive and compliant behavior on the ward. SWAP diagnoses were also predictive of patients' index offense (e.g., whether it was violent), whereas SCID–II diagnoses were not. The findings are clearly preliminary, but they provide some of the first data directly assessing the incremental validity of the SWAP–200 relative to a widely used PD instrument that relies substantially on patient self-report.

Potential Uses of Clinician-Report Data in Research on Psychopathology

Having established that clinicians can provide reliable and valid data, and that they can do so using instruments designed for experienced clinical observers, we now turn to the question of what might be gained by aggregating such data across clinicians. As an example, we briefly describe a study intended as a first step toward developing a classification of adolescent personality pathology (Westen et al., 2003). This same approach has proven useful in taxonomic work with other populations, such as eating disorders and adult PDs (see Westen & Harnden-Fischer, 2001; Westen & Shedler, 1999b).

A growing body of research over the last decade suggests that personality syndromes such as borderline PD are recognizable in adolescence (Bernstein, Cohen, Skodol, Bezirganian, & Brook, 1996; Grilo et al., 1998; Levy et al., 1999; Ludolph et al., 1990). To what extent Axis II of the *DSM–IV* represents an optimal way of classifying or diagnosing adolescent personality pathology is, however, unknown. The study described here used the adolescent version of the SWAP–200 Q sort, the SWAP–200–A. To develop the adolescent version of the instrument, we deleted, revised, and added items as appropriate based on the adolescent literature, the investigators' prior adolescent research and clinical experience, and consultation with senior adolescent clinicians who used the instrument to describe patients and then provided feedback on items that were ambiguous, necessary for describing their patient but missing from the item set, and so forth. As with the adult version, items were written in a manner close to the data (e.g., "Tends to run away from home" or "Has an exaggerated sense of self-importance"), and items requiring substantial inferences about internal processes were stated in simple language devoid of jargon (e.g., "Tends to blame others for own failures or shortcomings; tends to believe his/her problems are caused by external factors"). Participants in this study were the 294 psychologists and psychiatrists who participated in the CBCL study described above. Each clinician used the SWAP–200–A to describe a randomly selected adolescent patient (operationalized as "the last patient you saw last week before completing this form who meets study criteria"). Patients met inclusion criteria if they were between the ages of 14 and 18 and were being treated for "enduring maladaptive patterns of thought, feeling, motivation, or behavior," a definition of personality pathology we deliberately kept broad and nonrestrictive. We collected a stratified random sample of patients, stratifying by age and gender.

To identify naturally occurring diagnostic groupings, we used Q-factor analysis, a technique designed to identify clusters of patients who share common psychological features but that does not, like most other clustering techniques, assume mutually exclusive categories. This technique has been used successfully in studies of normal (e.g., Block, 1971; Caspi, 1998; Robins, John, Caspi, Moffitt, & Stouthamer-Loeber, 1996) and disordered (Westen & Shedler, 1999a, 1999b) personality. Q-factor analysis identified five clinically coherent, nonredundant diagnostic prototypes, which we labeled antisocial–psychopathic, emotionally dysregulated, avoidant–constricted, narcissistic, and histrionic, and one less severe personality *style*, a high-functioning internalizing style labeled inhibited self-critical. As with studies using the adult instrument, patients' PD scale scores (reflecting the degree of match between their 200-item profile and each of the empirically derived prototypes) showed predictable associations with measures of adaptive functioning (e.g., history of suicide attempts and arrests) as well as a range of other criterion variables relevant to construct validity.

Consider the empirically derived antisocial–psychopathic prototype. This prototype was characterized by items indicating a tendency to be rebellious or defiant toward authority figures; to express intense and inappropriate anger; to act impulsively; to blame others for one's own failures or shortcomings; to react to criticism with rage or humiliation; to be unreliable and irresponsible; to draw pleasure or self-esteem from being, or being seen as, "bad" or "tough"; to have emotions that spiral out of control; to seek thrills, novelty, and adventure; to break things or become physically assaultive when angry; to feel misunderstood, mistreated, or victimized; and to be unconcerned with the consequences of one's actions. This prototype closely resembles the construct of psychopathy in adults (Cleckley, 1941; Hare, Hart, & Harpur, 1991) as well as the more malignant, early onset forms of delinquent behavior identified by Moffitt and others (Moffitt, Caspi, Harrington, & Milne, 2002). High scores on this dimension predicted poor school performance; an arrest history; family history of alcohol abuse, illicit substance abuse, and criminality; and a history of physical abuse in childhood. These data suggest not only that clinicians can describe patients in ways that predict theoretically relevant criterion variables, but also that their personality descriptions can be aggregated statistically to generate constructs with theoretically meaningful correlates.

Q-factor analysis also generated diagnostic groupings that have not been identified previously using self-report and structured interview data but have consistently emerged in studies applying clinician-report data to adult samples (Shedler & Westen, 1998; Westen & Shedler, 1999b; Zittel & Westen, in press). Of particular relevance is the distinction between two kinds of adolescents who currently meet *DSM–IV* criteria for borderline PD but differ in substantial ways, just as they do in adult samples. Emotionally dysregulated adolescents are characterized by intense, distressing, poorly modulated emotions that spiral out of control and lead to desperate attempts to regulate them, such as self-mutilation and suicide attempts and gestures. Histrionic–borderline adolescents are characterized by dramatic, rather than primarily dysphoric, affect; classic histrionic traits, such as seductiveness and theatrics; and problematic attachment patterns, such as neediness, dependency, and rejection sensitivity. The identification of these two distinct personality constellations in multiple samples with both adolescents and adults suggests that the comorbidity of borderline, histrionic, and dependent PD observed in multiple studies using *DSM–III*, *–III–R*, and *–IV* criteria may be an artifact of overlapping diagnostic categories and criterion sets that do not adequately mirror the nature of personality pathology seen in clinical practice across a range of sites (e.g., outpatient, inpatient, school, forensic).

What is particularly worth noting here is that, in these studies, we were not interested in clinicians' implicit or explicit classification systems. The diagnostic distinctions described here emerged despite clinicians' familiarity with the *DSM–IV* diagnostic categories, even when we asked them to describe patients using the *DSM–IV* categories and regardless of whatever theoretical and classificatory beliefs or biases they may have professed. To put it another way, this research does not survey clinicians' opinions, any more than research using the BDI surveys patients' opinions about the nature or factor structure of depression. Rather, it asks clinicians to do what they should, theoretically, be able to do well—to observe phenomena in their domain of expertise, including phenomena that require considerable inference—and not what they should be unable to do well—to aggregate observations into intuitive categories or diagnoses by trying to intuit patterns of covariation over hundreds of cases across hundreds of often ill-defined variables. This is precisely what Meehl said clinicians could and could not do. The data appear to bear him out.

It is instructive, furthermore, to note precisely where participants in this and similar studies do and do not apply "clinical" judgment and, by extension, where they are and are not applying "actuarial" judgment. On the one hand, clinicians are not simply counting behaviors. The *DSM* has increasingly reduced the inferential demands on clinicians and research interviewers over successive editions to maximize reliability by eliminating or avoiding diagnostic criteria that are difficult to assess by self-report (e.g., imperviousness to consequences, a component of the psychopathy construct that is absent from the antisocial diagnosis). In contrast, the SWAP–200–A, like the adult version of the instrument, assumes a certain level of clinical sophistication, requiring clinicians not only to be able to indicate the presence or frequency of certain behaviors (e.g., self-mutilation, running away from home, losing jobs), but also to judge the extent to which patients regulate emotions in particular ways (e.g., "Tends to express aggression in passive and indirect ways; e.g., may make mistakes, procrastinate, forget, become sulky, etc."), view themselves and others in particular ways (e.g., "Appears unable to describe important others in a way that conveys a sense of who they are as people; descriptions of others come across as two-dimensional and lacking in richness"), and so forth. Clinicians using the adolescent version of the instrument also typically integrate information across data sources, such as parents and schools, in making judgments about individual items.

On the other hand, we do not ask clinicians to determine whether the patient crosses some arbitrary threshold for presence or absence of antisocial–psychopathic PD, whether the patient is likely to be suicidal in the next six months, or whether the patient is likely to get into further trouble with the law. Rather, we ask clinicians to describe their patient using 200 personality-descriptive statements, which assumes their capacity to observe and make inferences at a moderate level of generality. We then apply actuarial methods (in this case, a simple correlation coefficient) to gauge the extent to which the patient matches an empirical prototype of antisocial–psychopathic patients, patients who have made a suicide attempt in the six months following evaluation, patients who did or did not respond to cognitive–behavioral therapy or to Zoloft, patients who subsequently battered their spouse (Porcerelli, Cogan, & Hibbard, 2004), and so forth. In so doing, we transform valid clinical judgment into valid statistical prediction.

Summary

In the typical study comparing clinical and statistical prediction, the clinician's task is to integrate the available data (sometimes from interviews, sometimes from specific projective tests, sometimes from a single test, sometimes from combinations of these) and make a global judgment about the presence or absence of some phenomenon or predict the likelihood of some prior or future event. Often the criterion variable is something the clinician rarely or infrequently encounters and about which the clinician has no special knowledge or expertise. This represents a confluence of factors for which clinical prognostication is likely to be least valid. If one wants to know whether clinical experience confers any advantage in making broad diagnostic or prognostic judgments without benefit of statistic aggregation and without confounding informant effects with aggregation effects, one should compare informal predictions made by experienced clinicians with informal predictions by laypeople (Quadrant II vs. Quadrant IV of Figure 1), using a dependent variable for which clinical training and experience should confer expertise (e.g., diagnosing narcissism or psychosis following an hour-long interview).

The most important test of clinical judgment, however, would examine the incremental validity of quantified clinical inference using a psychometric instrument designed for that purpose (e.g., the SWAP–200) relative to quantified lay judgment using a well-validated self-report measure (e.g., the MMPI–II) in predicting clinically relevant dependent measures (e.g., informant reports, laboratory measures of implicit attentional biases, behavior outside the laboratory, prognosis, and treatment response). Such a study is the obvious next step in this program of research. Unfortunately, despite 50 years of research on an issue at the heart of clinical and personality psychology—and central to the validity of the scientific study of psychopathology, given that patient self-reports constitute the heart of diagnosis in virtually all clinical research—such data do not yet exist. What we *can* conclude at this point is that clinicians can provide valid and reliable data if we quantify their inferences using psychometric instruments designed for expert observers.

Conclusions and Implications

We conclude with one final complexity that bears on the broader way the clinical–statistical debate has been framed: the issue of whether any judgment can be dichotomously coded as *either* clinical or statistical (see also Holt, 1958).

When Statistical Inference Becomes Clinical Judgment

At one level, *all* judgment is ultimately clinical in Meehl's sense (i.e., informal and synthetic rather than actuarial). Except for trivial cases, such as assessment of biological sex, all observations in psychology, no matter how well quantified, inherently involve *some* informal aggregation over time and across situations by *someone*, whether an informant or a presumed expert interpreter of the data (or both).

For example, when a patient responds to an item on the BDI with a judgment about how well the item describes the way she thinks and feels, she is intuitively abstracting across space and time, assessing the intent of the questioner, comparing her current state to some implicit reference group (e.g., depressed people she knows or recollection of her own past experience), and so forth (Schwartz, 1999).[4] In asking informants to respond to questionnaire items, we are simply pushing informal, "clinical" aggregation back a step (or down a hierarchical level). Rather than answering a single broad question similar to those often asked of clinicians in studies of clinical prediction (e.g., "What is the likelihood that you will kill yourself in the next few months?"), we ask informants to aggregate observations and inferences in answering multiple, more specific questions (e.g., "How often do you think about suicide?" or "Do you often feel hopeless?"). We then aggregate their responses to multiple such questions and hence both maximize reliability of measurement and provide more potential predictor variables for a regression equation. Similarly, we do not typically ask people if they are introverts. Rather, we ask them multiple questions at a lower level of generality that allow us to make statistical inferences about the extent to which they are high or low on the latent construct of introversion.

The situation confronting a scientist analyzing data—or synthesizing the findings of a research literature—is no less "clinical" (i.e., informal, synthetic, and fallible) in Meehl's sense than the task confronting a clinician trying to formulate a case. Data require interpretation. This is why scientists often disagree. Factor analysis, a quintessentially actuarial procedure for deriving meaning from data, is probably best categorized in Meehl's terms as clinical aggregation of psychometric data: The factor analyst must decide which extraction methods and estimation procedures are the most appropriate for the sample, population, and constructs of interest; whether the available data support exploratory or confirmatory procedures; which factor solutions are the most theoretically coherent; and, most "clinically" of all, how to name the factors.

Clinical/informal judgment is equally pronounced when researchers attempt to synthesize a body of research. The practitioner of science in this instance is the expert prognosticator, who, through contemplation of the available data, must arrive at some synthetic judgment. This judgment typically takes the form of a hypothesis or network of hypotheses (a model) or an implicit or explicit set of predictions (e.g., of what studies are necessary to address key remaining issues). Quantitative methods of data aggregation are, in this case as in the situations we have been considering in this article, extremely helpful—hence the utility of meta-analysis. Ultimately, however, the practitioner of science must make nonquantitative, informal judgments (e.g., about the validity of particular studies or analyses) that are vulnerable to theory-driven confirmatory biases and other heuristics and biases documented years ago by Kuhn (1962) in his (empirical) examination of scientific practices across time and disciplines.

Researchers routinely question the conclusions of meta-analyses on the basis of their authors' judgments regarding inclusion criteria, methods of aggregating the studies, and so forth, in a way that suggests the "clinical" nature of interpretation of even the most quantitative data. Consider, for example, the judgments reached by different commentators on a recent meta-analysis of data bearing on the validity of Rorschach indices (Hiller, Rosenthal, Bornstein, Berry, & Brunell-Neuleib, 1999). The investigators drew two primary conclusions from their meta-analysis: that effect sizes for Rorschach variables tended to be comparable to those obtained using the MMPI and that the MMPI had a small advantage in predicting diagnoses, whereas Rorschach variables had a small advantage in predicting behavior. The published responses to this (nonpartisan) meta-analysis were decidedly partisan, with some commentators nodding approvingly at the study and others dismissing it with a plethora of post hoc methodological complaints, including (our favorite) that its authors (which included Robert Rosenthal, whose chapter on meta-analysis was published around the same time in the *Annual Review of Psychology*; Rosenthal & DiMatteo, 2000) did not understand the basics of meta-analytic technique (Garb, Wood, Nezworski, & Grove, 2001). Indeed, our (informal) dichotomous coding of the published responses to Hiller et

[4] Meehl (1954, p. 17) noted this issue in passing in his 1954 book.

al.'s meta-analysis (supportive vs. dismissive, coded 0/1) led to a regression equation with a multiple R of 1.0 based on a single predictor variable: the prior published attitudes of the article's first author toward the validity of Rorschach data. This is not far from the (more serious) multiple R reported in Luborsky and colleagues' (Luborsky et al., 1999) meta-analysis predicting outcome of randomized controlled trials of psychotherapy from investigator allegiance, which found that most of the time one can predict which treatment condition will show the strongest effect in psychotherapy research based on the investigator's belief in it.

Or consider a recent high-profile article on biases and errors among practitioners (Spence, Greenberg, Hodge, & Vieland, 2003). The headings of the paper convey the general point of the article: "Willingness to Establish Standards Without the Protections of Rigorous Testing," ". . . Practice Based on Myth Rather Than Evidence," and "The Unfortunate Development of a 'Cult of Personality'" (p. 1084). These are familiar themes in the literature on biases in clinical judgment. However, the practitioners to whom these authors were referring in this editorial in the *American Journal of Human Genetics* were not clinical psychologists but practitioners of *research in human genetics*, who, they argued, routinely reject grants "on the basis of myth," reject manuscripts "for failing to adhere to dogma," and launch huge projects "on the strength of personality cults" (p. 1084). Historians and sociologists of science have documented instance after instance of such errors and biases in "clinical" judgment across every scientific discipline studied (e.g., Barber, 1961).

The biases and heuristics characteristic of scientific judgment appear to us to differ little in kind from those confronting the clinician listening to complex material in a clinical hour, which may explain the paradox to many who knew Meehl, of his simultaneous belief in the importance of a scientific attitude and in the value of clinical interpretation of patients' associations.[5] Are clinicians uniquely vulnerable to confirmatory biases? The history of science can be viewed as the history of confirmatory biases. The eminent scientist Lord Kelvin declared Roentgen's discovery of X-rays to be an elaborate hoax. In psychology, researchers believed for half a century that mental events are nonexistent, epiphenomenal, or irrelevant for a scientific psychology, and they generated thousands of studies confirming their view. The persistence of serial processing models of cognition for 30 years and models and measures of attitudes that ignored implicit attitudes for 80 years strikes us as no less egregious than the cognitive errors widely attributed to clinicians.

Are clinicians uniquely vulnerable to illusory correlations (Chapman & Chapman, 1967; Garb, 1998)? One could make an equally strong case that the history of personality and clinical psychology is the history of illusory correlations, as researchers have routinely failed to partial out method variance (reliance on self-reports for both the predictor and criterion variables) and shared item content in correlational research (see Nicholls, Licht, & Pearl, 1982). For example, Clifton, Turkheimer, and Oltmanns (2003) recently found a high correlation between self-reported PD symptoms and interpersonal problems assessed by the IIP. This seems like a sensible finding, which would normally be taken at face value and published in our best journals. But the investigators went a step further, examining whether the correlations held when using IIP data aggregated across multiple informants. In fact, participants' self-reported PD symptoms explained little variance in the interpersonal problems others identified in them—or even in the PD symptoms others identified in them. The data were consistent with a more general finding from their program of research: Aggregated peer reports yield highly consistent portraits of an individual's personality, but for some negative traits, such as characteristics of PDs, they tend to be correlated only modestly with self-reports (Thomas et al., 2003).

Try as we might to eliminate subjectivity in science, we can never transcend the fact that the mind of the scientist, clinician, or informant is the source of much of what we know and what we think we know but is really error. For better or worse, in Meehl's sense of the term, we are all clinicians.

Implications

If we distinguish the process of data aggregation from the nature of the observer, we may arrive at a more nuanced view of the mind of the clinician (see also Westen & Weinberger, in press). On the one hand, 50 years after Meehl's classic treatise, the evidence is even more clear that informal methods of aggregating data are unlikely to predict behavior as well as formal, actuarial methods when a domain of research is sufficiently advanced as to permit identification and reliable measurement of key variables useful for prediction. And 50 years of research have given us a better understanding of the conditions under which clinicians need to show more circumspection in their speculations and prognostications and to be cognizant of a range of biases and heuristics that can affect both expert and lay inference (Dawes et al., 1989).

On the other hand, in a frequently forgotten passage of his 1954 book, Meehl (pp. 72–73) pointed to two circum-

[5] For the skeptical reader, we cite Meehl himself from "Why I Do Not Attend Case Conferences": "Psychologists who visit Minneapolis for the first time and drop in for a chat with me generally show clinical signs of mild psychic shock when they find a couch in my office and a picture of Sigmund Freud on the wall. Apparently one is not supposed to think or practice psychoanalytically if he understands something about philosophy of science, thinks that genes are important for psychology" (1973, p. 225). Speaking of himself in the third person, Meehl (1973) wrote, "It is well-known that he [Meehl] . . . considers the purely theoretical personality research of academic psychologists to be usually naive and unrealistic when the researcher is not a seasoned, practicing clinician. . . . He [Meehl] took the trouble to become a diplomate of ABPP although in his academic position this had little advantage either of economics or of status. When he was chairman of the psychology department he had a policy of not hiring faculty to teach courses in the clinical and personality area unless they were practitioners and either had the ABPP diploma or intended to get it" (p. 226).

stances in which clinical judgment can be indispensable to our field.[6] First, in the context of scientific discovery, in which we are framing hypotheses (or, we would add, drafting items for which clinical observation may be useful), immersion in a phenomenon by an experienced observer can be crucial for identifying relevant variables. Here collaboration between clinicians and researchers could substantially improve the quality of scientific research. Those of us whose professional lives are weighted toward research can spend only a fraction of the time clinicians spend in contact with patients to identify phenomena that may be crucial to test. Such phenomena are not always apparent when we have thoroughly structured participants' responses to the extent required in the context of scientific justification (hypothesis testing). The fact that clinicians took the existence of implicit associational networks as axiomatic a century before researchers came to a consensus about their existence should give us pause before dismissing the potential contribution of clinical observations to empirical psychology (see Weinberger, in press; Westen, 1988, 1998). We would add that clinical observation, though obviously less useful than controlled research for hypothesis testing, can contribute in one important respect in the context of justification: by providing disconfirming instances (or what philosophers sometimes call existence proofs). As Hume argued, if we conclude, based on observation of 99 swans, that all swans are white, we can never be certain that a black swan is not just around the corner. Clinical observation can be a wonderful black swan generator.

The second circumstance Meehl identified in which the cognitive activity of the clinician is essential is in the synthetic process of culling through the myriad things a patient says and does in any clinical hour to recognize potentially meaningful patterns. As we have argued, this process is no different in kind from the process of culling through one's own research data or through an entire research literature in an effort to separate signal from noise and to organize the data in a way that is scientifically useful. In both cases, we are dependent on the mind of the "practitioner," prone as it is to errors, heuristics, and motivated distortions. The more we can rely on statistical aggregation as a prosthesis for data integration, the more we are likely to reach valid conclusions. But ultimately, some imperfect ("clinical") mind must interpret and synthesize imperfect data into theories, models, or hypotheses. In the end, what Meehl called clinical aggregation may simply be another name for cognition, with all its potential for bias and error.

Perhaps we would do well to heed the seemingly disparate warnings of Hume, Bacon, Freud, and Meehl. From Hume (and later Kant) we learned that we cannot escape the subjectivity of the observer—that we will never see the world exactly as it is. From Bacon we learned that we must try anyway, and that scientific method is our best guide. From Freud (and later Kahneman and Tversky, Dawes, and others) we learned that our minds can play all kinds of tricks on us, and that systematic self-reflection, self-scrutiny, and knowledge about the biases to which we are prone are as essential for clinicians and scientists as for our patients. And from Meehl we learned that the scientific mind and the clinical mind can coexist, if ambivalently, in a single field—indeed, in a single person—and that the dialectic between the two may be essential for a scientific psychology.

[6] As Grove and Meehl (1996) eloquently put it, "Policymakers should not accept a practitioner's unsupported allegation that something works when the only warrant for this claim is purported clinical experience. Clinical experience is an invaluable source of ideas. It is also the only way a practitioner can acquire certain behavioral skills, such as how to ask questions of the client. It is not an adequate method for settling disputes between practitioners, because they can each appeal to their own clinical experience" (p. 319).

REFERENCES

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4–18 and 1991 Profile*. Burlington: University of Vermont, Department of Psychiatry.

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders* (1st ed.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychological Society Observer. (2003). In appreciation: Paul E. Meehl (1920–2003). *Observer, 12*, 13.

Barber, B. (1961). Resistance by scientists to scientific discovery. *Science, 134*, 596–602.

Basco, M. R., Bostic, J. Q., Davies, D., Rush, A. J., Witte, B., Hendrickse, W., et al. (2000). Methods to improve diagnostic accuracy in a community mental health setting. *American Journal of Psychiatry, 157*, 1599–1605.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory–II*. San Antonio, TX: Psychological Corporation.

Bernstein, D. P., Cohen, P., Skodol, A., Bezirganian, S., & Brook, J. S. (1996). Childhood antecedents of adolescent personality disorders. *American Journal of Psychiatry, 153*, 907–913.

Betan, E., Heim, A., Zittel, C., & Westen, D. (2004). *The structure of countertransference phenomena in psychotherapy: An empirical investigation.* Unpublished manuscript, Emory University.

Block, J. (1971). *Lives through time*. Berkeley, CA: Bancroft.

Block, J. (1978). *The Q-sort method in personality assessment and psychiatric research*. Palo Alto, CA: Consulting Psychologists Press.

Block, J. (1995). A contrarian view of the Five-Factor approach to personality descriptions. *Psychological Bulletin, 117*, 187–215.

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

Bradley, R., Hilsenroth, M., & Westen, D. (2003). *Validity of SWAP–200 personality diagnosis in an outpatient sample*. Unpublished manuscript, Emory University.

Brammer, R. (2002). Effects of experience and training on diagnostic accuracy. *Psychological Assessment, 14*, 110–113.

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

Caspi, A. (1998). Personality development across the life span. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social emotional, personality development* (pp. 311–388). New York: Wiley.

Cassidy, J., & Shaver, P. R. (1999). *Handbook of attachment: Theory, research, and clinical applications*. New York: Guilford Press.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72*, 193–204.

Cleckley, H. (1941). *The mask of sanity*. St. Louis, MO: Mosby.

Clifton, A., Turkheimer, E., & Oltmanns, T. F. (2003). *Self and peer perspectives on pathological personality traits and interpersonal problems*. Unpublished manuscript, University of Virginia.

Colvin, R., Block, J., & Funder, D. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology, 68*, 1152–1162.

Cousineau, T. M. (1997). *Psychological predictors of health service utilization in college students*. Unpublished doctoral dissertation, Adelphi University.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1964). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*, 137–163.

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

Dozier, M., & Kobak, R. (1992). Psychophysiology in attachment interviews: Converging evidence for deactivating strategies. *Child Development, 63*, 1473–1480.

Dutra, L., Campbell, L., & Westen, D. (2004). Quantifying clinical judgment in the assessment of adolescent psychopathology: Reliability, validity, and factor structure of the Child Behavior Checklist for Clinician-Report. *Journal of Clinical Psychology, 60*, 65–85.

Epstein, S. (1986). Does aggregation produce spuriously high estimates of behavior stability? *Journal of Personality and Social Psychology, 50*, 1199–1210.

Epstein, S. (1992). Coping ability, negative self-evaluation, and overgeneralization: Experiment and theory. *Journal of Personality and Social Psychology, 62*, 826–836.

Fiedler, E., Oltmanns, T., & Turkheimer, E. (in press). Traits associated with personality disorders and adjustment to military life: Predictive validity of self and peer reports. *Military Psychology*.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997). *Structured Clinical Interview for DSM–IV personality disorders* (SCID–II). Washington, DC: American Psychiatric Press.

Fonagy, P., Steele, H., & Steele, M. (1991). Maternal representations of attachment during pregnancy predict the organization of infant–mother attachment at one year of age. *Child Development, 62*, 891–905.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.

Garb, H. N., Wood, J. M., Nezworski, M. T., & Grove, W. M. (2001). Toward a resolution of the Rorschach controversy. *Psychological Assessment, 13*, 433–448.

Gilbert, D. T., & Ebert, J. E. J. (2002). Decisions and revisions: The affective forecasting of changeable outcomes. *Journal of Personality and Social Psychology, 82*, 502–514.

Goldberg, L. R. (1991). Human mind versus regression equation: Five contrasts. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul E. Meehl* (Vol. 1, pp. 173–184). Minneapolis: University of Minnesota Press.

Greenbaum, P. E., & Dedrick, R. F. (1998). Hierarchical confirmatory factor analysis of the Child Behavior Checklist/4–18. *Psychological Assessment, 10*, 149–155.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress. *Psychological Review, 93*, 216–229.

Grilo, C. M., McGlashan, T. H., Quinlan, D. M., Walker, M., Greenfeld, D., & Edell, W. (1998). Frequency of personality disorders in two age cohorts of psychiatry inpatients. *American Journal of Psychiatry, 155*, 140–142.

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.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19–30.

Hare, R. D. (1998). Psychopaths and their nature: Implications for the mental health and criminal justice systems. In T. Millon & E. Simonsen (Eds.), *Psychopathy: Antisocial, criminal, and violent behavior* (pp. 188–212). New York: Guilford Press.

Hare, R. D., Harpur, T. J., Hakstian, A. R., Forth, A. E., Hart, S. D., & Newman, J. P. (1990). The revised Psychopathy Checklist: Reliability and factor structure. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2*, 338–341.

Hare, R. D., Hart, S. D., & Harpur, T. J. (1991). Psychopathy and the *DSM–IV* criteria for antisocial personality disorder. *Journal of Abnormal Psychology, 100*, 391–398.

Harkness, A. R., Tellegen, A., & Waller, N. (1995). Differential convergence of self-report and informant data for Multidimensional Personality Questionnaire traits: Implications for the construct of negative emotionality. *Journal of Personality Assessment, 64*, 185–204.

Heim, A., & Westen, D. (2002). *Subclinical Cognitive Disturbance Inventory*. Unpublished manual, Emory University. Available from www.psychsystems.net/lab

Hiller, J. B., Rosenthal, R., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (1999). A comparative meta-analysis of Rorschach and MMPI validity. *Psychological Assessment, 11*, 278–296.

Holt, R. (1958). Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology, 56*, 1–12.

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureno, G., & Billasenor. (1988). Inventory of Interpersonal Problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology, 56*, 885–892.

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology, 66*, 206–219.

Johnston, M. H., & Holzman, P. S. (1979). *Assessing schizophrenic thinking*. San Francisco: Jossey-Bass.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–25l.

Kahneman, D., & Tversky, A. (2000). *Choice, values, and frames*. New York: Cambridge University Press.

Klonsky, E. D., Oltmanns, T. F., & Turkheimer, E. (2002). Informant-reports of personality disorder: Relation to self-reports and future research directions. *Clinical Psychology: Science & Practice, 9*, 300–311.

Kranzler, H., Kadden, R., Babor, T., Tennen, H., & Rounsaville, B. (1996). Validity of the SCID in substance abuse patients. *Addiction, 91*, 859–868.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480–498.

Levy, K. N., Becker, D. F., Grilo, C. M., Mattanah, J., Garnet, K. E., Quinlan, D. M., et al. (1999). Concurrent and predictive validity of the personality disorder diagnosis in adolescent patients. *American Journal of Psychiatry, 156*, 1522–1528.

Lewicki, P. (1986). Processing information about covariations that cannot be articulated. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 135–146.

Loevinger, J., & Wessler, R. (1970). *Measuring ego development: Construction and use of a Sentence Completion Test* (Vol. 1). San Francisco: Jossey-Bass.

Lorenz, A. R., & Newman, J. P. (2002). Utilization of emotion cues in male and female offenders with antisocial personality disorder: Results from a lexical decision task. *Journal of Abnormal Psychology, 111*, 513–516.

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., et al. (1999). The researcher's own therapy allegiances: A "wild card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice, 6*, 95–106.

Ludolph, P., Westen, D., Misle, B., Jackson, A., Wixom, J., & Wiss, F. C. (1990). The borderline diagnosis in adolescents: Symptoms and developmental history. *American Journal of Psychiatry, 147*, 470–476.

Main, M., Kaplan, N., & Cassidy, J. (1985). Security in infancy, childhood, and adulthood: A move to the level of representation. *Monographs of the Society for Research in Child Development* 50(1–2, Serial No. 209).

Marin-Avellan, L., McGauley, G., Campbell, C., & Fonagy, P. (2004, February). *Using the SWAP–200 in a personality-disordered forensic population: Is it valid, reliable and useful?* Paper presented at the annual Conference of the British and Irish Group for the Study of Personality Disorders, Cardiff, UK.

McAdams, D. (1992). The Five-Factor model in personality: A critical appraisal. *Journal of Personality, 60,* 329–361.

McClelland, D. C. (1985). *Human motivation*. Glenview, IL: Scott, Foresman.

McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review, 96*, 690–702.

McFall, R. (1991). Manifesto for a science of clinical psychology. *The Clinical Psychologist, 44,* 75–88.

McReynolds, P. (1987). Lightner Witmer: Little-known founder of clinical psychology. *American Psychologist, 42,* 849–858.

Meehl, P. E. (1954). *Clinical vs. statistical prediction*. Minneapolis: University of Minnesota Press.

Meehl, P. (1959). Some ruminations on the validation of clinical procedures. *Canadian Journal of Psychology, 13*, 106–128.

Meehl, P. E. (1960). The cognitive activity of the clinician. *American Psychologist, 15*, 19–27.

Meehl, P. (1973). Why I do not attend case conferences. In P. Meehl (Ed.), *Psychodiagnostics: Selected papers* (pp. 225–302). New York: Norton.

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.

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

Moffitt, T. E., Caspi, A., Harrington, H., & Milne, B. J. (2002). Males on the life-course-persistent and adolescence-limited antisocial pathways: Follow-up at age 26 years. *Development and Psychopathology, 14*, 179–207.

Morey, L. C. (1991). *The Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Nicholls, J. G., Licht, B. G., & Pearl, R. A. (1982). Some dangers of using personality questionnaires to study personality. *Psychological Bulletin, 92*, 572–580.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Novotny, C., Eddy, K., & Westen, D. (2004). *Impulsivity in eating disorders treated in the community*. Unpublished manuscript, Emory University.

Paulhus, D. L., Fridhandler, B., & Hayes, S. (1997). Psychological defense: Contemporary theory and research. In R. Hogan, J. Johnson, & S. Briggs, *Handbook of personality psychology* (pp. 543–579). New York: Academic Press.

Perry, J. C. (1992). Problems and considerations in the valid assessment of personality disorders. *American Journal of Psychiatry, 149*, 1645–1653.

Pilkonis, P. A., Heape, C. L., Proietti, J. M., Clark, S. W., McDavid, J. D., & Pitts, T. E. (1995). The reliability and validity of two structured diagnostic interviews for personality disorders. *Archives of General Psychiatry, 52*, 1025–1033.

Pilkonis, P. A., Heape, C. L., Ruddy, J., & Serrao, P. (1991). Validity in the diagnosis of personality disorders: The use of the LEAD standard. *Psychological Assessment, 31*, 46–54.

Porcerelli, J. H., Cogan, R., & Hibbard, S. (2004). Personality characteristics of partner violent men: A Q-sort approach. *Journal of Personality Disorders, 18*, 151–162.

Ready, R. E., Watson, D., & Clark, L. A. (2002). Psychiatric patient- and informant-reported personality: Predicting concurrent and future behavior. *Assessment, 9*, 361–372.

Reber, A. (1992). The cognitive unconscious: An evolutionary perspective. *Consciousness and Cognition, 1*, 93–133.

Robins, R. W., John, O., Caspi, A., Moffitt, T. E., & Stouthamer-Loeber, M. (1996). Resilient, overcontrolled, and undercontrolled boys: Three replicable personality types. *Journal of Personality and Social Psychology, 70*, 157–171.

Rosenthal, R., & DiMatteo, M. R. (2000). Meta analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59–82.

Rosnow, R. L., & Rosenthal, R. (1991). If you're looking at the cell means, you're not looking at only the interaction (unless all main effects are zero). *American Psychologist, 110*, 574–576.

Rubin, D. B., Wallace, W., & Houston, B. (1993). The beginnings of expertise for ballads. *Cognitive Science, 17*, 435–462.

Russ, E., Heim, A., & Westen, D. (2003). Parental bonding and personality pathology assessed by clinician report. *Journal of Personality Disorders, 17,* 522–536.

Sarbin, T. R. (1962). The present status of the clinical–statistical prediction problem. *Anthropology and Medicine, 10*, 315–323.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*, 178–200.

Schwartz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93–105.

Segal, D. L., Corcoran, J., & Coughlin, A. (2002). Diagnosis, differential diagnosis, and the SCID. In M. Hersen & L. K. Porzelius (Eds.), *Diagnosis, conceptualization, and treatment planning for adults: A step-by-step guide* (pp. 13–34). Mahwah, NJ: Erlbaum.

Shedler, J., Mayman, M., & Manis, M. (1993). The illusion of mental health. *American Psychology, 48*, 1117–1131.

Shedler, J., Mayman, M., & Manis, M. (1994). More illusions. *American Psychologist, 49*, 974–976.

Shedler, J., & Westen, D. (1998). Refining the measurement of Axis II: A Q-sort procedure for assessing personality pathology. *Assessment, 5*, 333–353.

Shedler, J., & Westen, D. (2004). Refining *DSM–IV* personality disorder diagnosis: Integrating science and practice. *American Journal of Psychiatry, 161*, 1–16.

Shedler, J., & Westen, D. (in press). Dimensions of personality pathology: An alternative to the Five Factor Model. *American Journal of Psychiatry*.

Smith, C. P., Atkinson, J. W., McClelland, D. C., & Veroff, J. (Eds.). (1992). *Motivation and personality: Handbook of thematic content analysis*. New York: Cambridge University Press.

Spence, M. A., Greenberg, D. A., Hodge, S. E., & Vieland, V. J. (2003). The emperor's new methods. *American Journal of Human Genetics, 72*, 1084–1087.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24*, 399–411.

Spitzer, R. L., Endicott, J., & Robins, E. (1975). Clinical criteria for psychiatric diagnosis and *DSM–III*. *American Journal of Psychiatry, 132*, 1187–1192.

Stricker, G., & Trierweiler, S. J. (1995). The local clinical scientist: A bridge between science and practice. *American Psychologist, 50*, 995–1002.

Strube, M. J. (2000). Reliability and generalizability theory. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding MORE multivariate statistics* (pp. 23–66). Washington, DC: American Psychological Association.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology, 23*, 457–482.

Tavris, C. (2003). Mind games: Psychological warfare between therapists and scientists. *Chronicle of Higher Education, 49*, B47.

Thomas, C., Turkheimer, E., & Oltmanns, T. F. (2003). Factorial structure of pathological personality as evaluated by peers. *Journal of Abnormal Psychology, 112*, 81–91.

Weinberger, J. (in press). *The rediscovery of the unconscious*. New York: Guilford Press.

Westen, D. (1988). Official and unofficial data. *New Ideas in Psychology, 6*, 323–331.

Westen, D. (1995). A clinical–empirical model of personality: Life after the Mischelian ice age and the NEO-lithic era. *Journal of Personality, 63*, 495–524.

Westen, D. (1996). A model and a method for uncovering the nomothetic from the idiographic: An alternative to the Five-Factor Model? *Journal of Research in Personality, 30*, 400–413.

Westen, D. (1997). Divergences between clinical and research methods for assessing personality disorders: Implications for research and the evolution of Axis II. *American Journal of Psychiatry, 154*, 895–903.

Westen, D. (1998). The scientific legacy of Sigmund Freud: Toward a psychodynamically informed psychological science. *Psychological Bulletin, 124*, 333–371.

Westen, D. (2002). *Clinical Diagnostic Interview*. Unpublished manual, Emory University. Available from www.psychsystems.net/lab

Westen, D., Feit, A., & Zittel, C. (1999). Methodological issues in research using projective techniques. In P. C. Kendall, J. N. Butcher, & G. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (2nd ed., pp. 224–240). New York: Wiley.

Westen, D., & Harnden-Fischer, J. (2001). Personality profiles in eating disorders: Rethinking the distinction between Axis I and Axis II. *American Journal of Psychiatry, 165*, 547–562.

Westen, D., & Muderrisoglu, S. (2003). Reliability and validity of personality disorder assessment using a systematic clinical interview: Evaluating an alternative to structured interviews. *Journal of Personality Disorders, 17*, 350–368.

Westen, D., Muderrisoglu, S., Fowler, C., Shedler, J., & Koren, D. (1997). Affect regulation and affective experience: Individual differences, group differences, and measurement using a Q-sort procedure. *Journal of Consulting and Clinical Psychology, 65*, 429–439.

Westen, D., Novotny, C., & Thompson-Brenner, H. (2004). The empirical status of empirically supported therapies: Assumptions, methods, and findings. *Psychological Bulletin, 130*, 631–663.

Westen, D., & Shedler, J. (1999a). Revising and assessing Axis II, Part 1: Developing a clinically and empirically valid assessment method. *American Journal of Psychiatry, 156*, 258–272.

Westen, D., & Shedler, J. (1999b). Revising and assessing Axis II, Part 2: Toward an empirically based and clinically useful classification of personality disorders. *American Journal of Psychiatry, 156*, 273–285.

Westen, D., Shedler, J., Durrett, C., Glass, S., & Martens, A. (2003). Personality diagnosis in adolescence: *DSM–IV* Axis II diagnoses and an empirically derived alternative. *American Journal of Psychiatry, 160*, 952–966.

Westen, D., & Weinberger, J. (in press). In praise of clinical judgment: Meehl's forgotten legacy. *Journal of Clinical Psychology.*

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

Wilberg, T., Dammen, T., & Friis, S. (2000). Comparing personality diagnostic questionnaire-4+ with longitudinal, expert, all data (LEAD) standard diagnoses in a sample with a high prevalence of Axis I and Axis II disorders. *Comprehensive Psychiatry, 41*, 295–302.

Wilkinson-Ryan, T., & Westen, D. (2000). Identity disturbance in borderline personality disorder: An empirical investigation. *American Journal of Psychiatry, 157*, 528–541.

Williams, J. M., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120*, 3–24.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*, 101–126.

Wood, J. M., Garb, H. N., Lilienfeld, S. O., & Nezworski, M. (2002). Clinical assessment. *Annual Review of Psychology, 53*, 519–543.

Zittel, C., & Westen, D. (in press). Borderline personality disorder as seen in clinical practice: Implications for *DSM–V*. *American Journal of Psychiatry*.