In Praise of Clinical Judgment: Meehl's Forgotten Legacy

▼

Drew Westen
Emory University

▼

Joel Weinberger
Derner Institute of Advanced Psychological Studies, Adelphi University

Although Paul E. Meehl demonstrated the limits of informal aggregation of data and prognostication by presumed experts, he remained convinced that clinical experience confers expertise of some kind. The authors explore this forgotten side of Meehl's legacy by reconsidering the validity of clinical judgment in its natural context, everyday clinical work. Three domains central to clinical practice are examined: diagnosis, interpretation of meaning, and intervention. It is argued that a more sanguine picture of clinical expertise emerges when the focus shifts from prediction at high levels of inference to (a) judgments at a moderate level of inference, (b) contexts for which clinical training and experience are likely to confer expertise, and (c) conditions that optimize the expression of that expertise (e.g., use of instruments designed for expert observers). The authors conclude by examining domains in which clinical judgment could prove useful in knowledge generation (e.g., hypothesis generation, identification of falsifying instances, item development). © 2005 Wiley Periodicals, Inc. J Clin Psychol 61: 1257–1276, 2005.

Keywords: Paul E. Meehl; clinical prediction; statistical prediction; clinical judgment; clincial decision making; diagnosis

Few of us will live long enough, or write anything of sufficient profundity, to see our work cited or assigned to graduate students 50 years later. Yet Meehl's work from a half-century ago remains central to the canon of our discipline. At the heart of this work was his book on clinical and statistical prediction (Meehl, 1954/1996), in which he first

Correspondence concerning this article should be addressed to: Drew Westen, Department of Psychology, Emory University, 532 Kilgo Circle, Atlanta, Georgia 30322; e-mail: dwesten@emory.edu

JOURNAL OF CLINICAL PSYCHOLOGY, Vol. 61(10), 1257–1276 (2005) © 2005 Wiley Periodicals, Inc.
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jclp.20181

demonstrated the relative superiority of actuarial to informal modes of aggregating data, and his blistering critique of sloppy clinical thinking in "Why I Do Not Attend Case Conferences" (Meehl, 1973). This one–two punch, combining a devastating quantitative jab with a qualitative right hook, staggered the emerging field of clinical psychology, and stands to this day as a testament to the rightful place of hubris among the deadly sins. 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), but the weight of the evidence remains the same as it was a half-century ago: Those who believe they can "beat the odds" of a well-developed, well-validated formula would do well to keep their wagers small (Grove, Zald, Lebow, Snitz, & Nelson, 2000).

Meehl's work on clinical–statistical prediction is often understood as a condemnation of clinicians and their characteristic modes of thought and belief. However, from the start, Meehl and his collaborators (Dawes et al., 1989; Grove & Meehl, 1996; Grove et al., 2000) distinguished what they meant by *clinical* (an informal, subjective, nonquantitative mode of aggregating observations to make predictions) from its broader connotation (pertaining to the judgments, inferences, observations, beliefs, or practices of clinicians; Westen & Weinberger, 2004). In fact, Meehl was driven to write his book on clinical–statistical prediction by what he described in its preface as a conflict confronting anyone who practices both research and psychotherapy, between the subjective certainty that clinical experience, like other forms of experience, surely must confer expertise, and the disappointing findings on the reliability and validity of diagnostic judgments and prognostications by purported experts.

Meehl argued persuasively that informal aggregation of data leaves clinicians (and other experts) open to the same kinds of judgmental biases and heuristics subsequently identified by Kahneman and Tversky (1973), Nisbett and Ross (1980), and others as characteristic of everyday cognition. Yet he never abandoned his belief in clinical knowledge, judgment, or expertise. For example, he patiently awaited data supporting clinically based hypotheses about the pervasiveness, complexity, and patterning of unconscious processes as he practiced psychoanalysis into his 80s (see Meehl, 1983). In Meehl's own words:

> [P]sychologists 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 . . . [M]y local psychonomic brethren find it odd that I should be seriously interested in the interpretation of dreams. (1973, pp. 225–226)

Meehl had a strong distrust for clinical theories and research that did not reflect immersion in clinical work and (referring to himself in the third person), "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).

Our goal in this article is to revisit this largely forgotten side of Meehl's attitude toward clinicians and the clinical enterprise, examining the circumstances under which clinicians can make valid inferences. We are convinced, as was Meehl, that informal, subjective prognostication is not the forté of clinicians, for precisely the reasons he and others have elucidated: Optimal prediction requires a standard set of predictor variables and regression weights iteratively selected and cross-validated across multiple data sets, selected for their validity in predicting known outcomes, and aggregated statistically.

However, except in the courtroom (where prognostications of presumed experts can be notoriously unreliable) and in the mass media (where clinicians with a passion for the limelight and limited self-control often opine about the diapers of snipers), such prognostications are not (and should not be) the stuff of everyday clinical practice and hence may not be the decisive test of clinical judgment.

To put it another way, the clinical–statistical prediction debate in psychology generally assumes a particular goal of clinical observation: prediction of behavior.[1] (It has also historically included psychiatric diagnosis, although, as we argue below, prognostication and diagnosis may have very different properties vis-à-vis the clinical-statistical debate.) In one sense, this is eminently reasonable, given that clinicians make predictions all the time, at least implicitly, as when they choose a particular avenue for intervening at a given point in a session with a particular patient. However, except in specific medical–legal situations (predicting suicidality or homicidally) or subspecialties (e.g., forensics), clinical training is typically not devoted to prediction, and clinical experience does not confer many benefits in predicting behavior, particularly without the kind of systematic feedback that is essential for self-correction (and built into regression equations). Much of clinical training is devoted instead to what clinicians need to know to do their work: how to diagnose personality and psychopathology, how to understand what their patients are doing or saying (or not saying), and how to intervene to help alleviate their patients' problems.

In this article, we address the validity of clinical judgment in its natural context, everyday clinical work, examining three domains: diagnosis, interpretation of meaning, and intervention. We focus on clinical judgment at moderate levels of inference (e.g., whether a patient is rejection sensitive, irresponsible, remorseless, or has difficulty imagining what other people feel) rather than on broad, unstructured prognostications (e.g., whether the person is likely to succeed in the army or to become manic at some point in the future), because this is the level of inference most germane to everyday clinical work and hence the most appropriate test of the validity of clinical judgment. Prognostications at higher levels of inference inherently confound two variables: the nature of the observer (clinician vs. lay) and the way the observer is trying to answer the question (statistically or intuitively) (Westen & Weinberger, 2004). If we know that statistical aggregation is generally superior to informal, subjective guesswork in predicting relatively broad or distal outcomes, we will readily conclude that clinicians lack expertise if predicting such outcomes is the measure of their expertise. If we want to see whether experienced practitioners accrue knowledge by virtue of their training and experience, there we would do better to focus on indices sensitive to the kind of expertise we might expect them to have (and to discourage them from making the kinds of predictions neither experts nor laypeople are likely to be able to make with any accuracy). We conclude by briefly revisiting the question of what clinicians might be able to contribute to knowledge generation (theory and research—that is, whether clinical experience might be not only clinically but scientifically useful.

Clinical Judgment and Clinical Diagnosis

When Meehl (1954/1996) first penned his critique of clinical prediction, the scope of his critique included clinical diagnosis, which had proven hopelessly unreliable. In retrospect, diagnosis using the *DSM-I* and *DSM-II* (*Diagnostic and Statistical Manual of*

[1]More broadly, "clinical prediction" in Meehl's use of the term could include predictions made by other experts, such as stock market forecasts by mutual fund managers.

Mental Disorders, first and second editions; American Psychiatric Association, 1952, 1968) could not have been otherwise. The first two editions of the *DSM* were coding manuals without coding rules, supplying broad diagnostic categories with no specific criteria or decision rules for adjudicating borderline cases. Like subsequent editions of the *DSM*, these early editions of the diagnostic manual also asked clinicians to make binary (present/absent) decisions about diagnostic variables (global diagnoses in *DSM-I* through *DSM-II*, individual diagnostic criteria in *DSM-III* through *DSM-IV*) that are more often than not continuously distributed in nature (see Westen & Shedler, 1999a; Widiger & Clark, 2000).

With the advent of *DSM-III* (1980), replete with highly specific, operationalizable diagnostic criteria, research surged ahead of practice in diagnostic reliability. Despite the ubiquitous focus in clinical training on *DSM* (or International Classification of Diseases; ICD) categories and criteria, clinicians still frequently do not know the criteria for particular disorders, do not use the diagnostic algorithms spelled out in the manual, and rely only loosely on the manual to make diagnoses (Garb, 1998; Jampala, Sierles, & Taylor, 1988). These facts have led many researchers to conclude that clinicians should abandon their preferred clinical interviewing and diagnostic practices in favor of structured research interviews so that they can make proper diagnoses (e.g., Basco et al., 2000; Segal, Corcoran, & Coughlin, 2002).

That the use of structured interviews would improve reliability of diagnosis in everyday practice is manifestly true. What is seldom recognized, however, is that this is true by definition. To make a reliable *DSM-IV* diagnosis, one must make hundreds of highly specific, often arbitrary decisions (one for each criterion for each disorder). This can only be done by systematically inquiring about every sign and symptom in specific ways required to decide whether the patient meets what are usually arbitrary cut-offs for "caseness" (e.g., whether the patient has binged and purged twice per week on average rather than once per week). It is difficult to imagine how one could make even moderately reliable diagnostic judgments using the diagnostic algorithms specified in the *DSM-IV* without administering a structured interview.

The fine-grained distinctions required to obtain reliability using the *DSM-IV* may be essential for researchers trying to identify homogeneous patient groups, but such distinctions are arguably of little relevance to clinicians, for whom the nature and goals of classification overlap with, but are not identical to, those of researchers. From a strictly empirical perspective, we are aware of no evidence that patients who fall just below or just above threshold for the categorical diagnosis of any *DSM-IV* diagnosis respond differently to any form of treatment, have different etiologies, or differ in any other important respect. The *DSM-III*, which first imposed the complex diagnostic decision rules characteristic of subsequent editions of the manual, was a direct outgrowth of the *Research* Diagnostic Criteria (RDC; Spitzer, Endicott, & Robins, 1978; emphasis added). The purpose of the RDC was to operationalize diagnoses so that researchers could know, at least by convention, that when they were studying patients with a particular diagnosis (e.g., major depression) in one hospital they would likely obtain similar data in another hospital. The hope was that this procedure would ultimately lead to better understanding of psychiatric disorders and better treatment.

In this sense, the RDC and its descendants, *DSM-III* through *DSM-IV*, have been a resounding success. However, from a clinical point of view, whether a patient with an eating disorder purges (e.g., through self-induced vomiting) once a week, twice a week, or sometimes once and sometimes twice is often of little consequence. Knowing that she has been purging once or twice a week for the last 2 years is usually good enough, and trying to pin her down on exactly how many times she has purged per week for a precise number of months so that one can decide whether she "really" has bulimia nervosa is

arguably not a responsible use of clinical time. The more important diagnostic questions, from a clinical standpoint, often center on issues about which *DSM-IV* diagnosis is silent, such as the circumstances that elicit purging in this particular patient, the extent to which the patient can control her tendency to purge, the extent to which she can take her purging or her beliefs about her body as an object of thought (rather than assuming them to be true), the extent to which her purging has become medically dangerous, the extent to which it is related to other forms of impulse dysregulation to which she may be prone, and so forth (see Westen & Bradley, 2005; Westen, Heim, Morrison, Patterson, & Campbell, 2002).

To what extent clinicians can reliably answer the kinds of diagnostic questions that are most useful in clinical practice is unknown. As a field, we have paid surprisingly little empirical attention to identifying the kinds of information clinicians might find useful in guiding interventions and in ways to optimize reliability and validity of those judgments. However, an emerging body of evidence using quantified clinician-report measures, designed using the same psychometric principles Meehl and others pioneered and personality and clinical researchers have applied successfully for 50 years to self-reports, suggests that clinicians can answer many kinds of clinically meaningful diagnostic questions with considerable reliability and validity if given the means for quantifying their observations (Dutra, Campbell, & Westen, 2004; Westen & Weinberger, 2004). For example, clinicians can make highly reliable diagnostic judgments about complex personality patterns using a 200-item Q-sort (e.g., Westen & Muderrisoglu, 2003; Westen & Shedler, 1999a, 1999b; Westen, Shedler, Durrett, Glass, & Martens, 2003). A Q-sort is a simple ranking procedure whereby an observer, in this case a clinician, rank-orders a set of items by sorting them into piles based on the extent to which they are descriptive of a target, in this case, a patient (see Block, 1978). Interestingly, Meehl himself mused about such a possibility: "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).

A Q-sort procedure might be useful in clinical practice when clinicians find themselves confronting a confusing diagnostic picture or when they must answer a precise, clinically or socially important question, such as whether the patient is a high suicide risk or a likely recidivist for a particular offense, for which researchers could develop empirical prototypes or algorithms that could allow more accurate statistical prediction. Clinicians' responses to Q-sort items written at a level likely to tap clinical expertise and amenable to statistical aggregation could be a powerful tool in such situations.

In everyday clinical practice, however, clinicians may be able to make reasonably reliable and valid diagnostic judgments at a moderate level of generality using a much simpler procedure. Westen, Shedler, and colleagues have been experimenting with a simple prototype-matching approach to diagnosis of both Axis I and Axis II syndromes that, for routine clinical purposes, might serve as a proxy for more precise Q-sort diagnoses and might represent a compelling alternative to the complex decision rules embodied in the *DSM-IV* (Westen & Bradley, 2005; Westen et al., 2002; Westen & Shedler, 2000). Using this approach, the clinician compares the patient's clinical presentation with a set of paragraph-long syndrome descriptions and makes a 1–5 rating of degree of match to each prototype, where a rating of 1 means "no match" and a rating of 5 means "prototypical case" (Figure 1). Rather than counting criteria assessed independently of one another, the clinician's task is to judge the goodness of fit between the prototype *taken as a whole*, or as a gestalt, and the patient's symptom picture. This approach has the advantage of providing both continuous and categorical diagnostic judgments: A rating of 4 or 5 constitutes, for purposes of communication among clinicians and researchers, a

5	Very good match (patient *exemplifies* this disorder; prototypical case)	Diagnosis
4	Good match (patient *has* this disorder; diagnosis applies)	
3	Moderate match (patient has *significant features* of this disorder)	Features
2	Slight match (patient has minor features of this disorder)	
1	Little or no match (description does not apply)	

Patients who match this prototype tend to be deceitful, and tend to lie and mislead others. They take advantage of others, have minimal investment in moral values, and appear to experience no remorse for harm or injury caused to others. They tend to manipulate others' emotions to get what they want; to be unconcerned with the consequences of their actions, appearing to feel immune or invulnerable; and to show reckless disregard for the rights, property, or safety of others. They have little empathy and seem unable to understand or respond to others' needs and feelings unless they coincide with their own. Individuals who match this prototype tend to act impulsively, without regard for consequences; to be unreliable and irresponsible (e.g., failing to meet work obligations or honor financial commitments); to engage in unlawful or criminal behavior; and to abuse alcohol. They tend to be angry or hostile; to get into power struggles; and to gain pleasure or satisfaction by being sadistic or aggressive toward others. They tend to blame others for their own failures or shortcomings and believe that their problems are caused entirely by external factors. They have little insight into their own motives, behavior, etc. They may repeatedly convince others of their commitment to change but then revert to previous maladaptive behavior, often convincing others that "this time is really different."

Figure 1. Prototype description for empirically derived antisocial personality disorder.

categorical diagnosis, and a rating of 3 denotes "features" of the disorder. For patients who receive a rating of 3 or higher on a given diagnosis, the clinician would make secondary ratings on such variables as age of onset and severity of symptom constellations that have proven empirically predictive or clinically useful (e.g., depressed mood, vegetative signs, suicidality) or that help distinguish genuinely taxonic from nontaxonic cases (see Meehl, 1995a).

Westen and colleagues have now applied this approach in several studies and are obtaining a consistent portrait of the reliability and validity of experienced doctoral-level clinicians' diagnostic judgments when using a diagnostic approach that makes better use of clinical judgment. Across a range of disorders, including personality, eating, mood, and anxiety disorders, prototype diagnosis appears to perform as well or better in *prediction* than application of *DSM-IV* diagnostic algorithms; reduces artifactual comorbidity; and is rated by clinicians as substantially more clinically useful, relevant, and efficient.

In one set of studies (Westen, Shedler, & Bradley, 2004), clinicians rated the extent to which a randomly selected patient in their care resembled single-paragraph prototype descriptions of the Axis II Cluster B (dramatic, erratic) personality disorders (antisocial, borderline, histrionic, and narcissistic) or prototypes of the same disorders derived empirically using Q-factor analysis in a prior study. Clinicians also completed a checklist of all the criteria for each of the *DSM-IV* Axis II disorders, which allowed the investigators to generate both categorical diagnoses and dimensional diagnoses (number of criteria met per disorder) for each disorder by using *DSM-IV* algorithms for diagnoses (e.g., presence of five of nine criteria).

Both sets of prototypes yielded substantially reduced estimates of comorbidity as well as higher ratings of clinical utility than diagnosis using *DSM-IV* decision rules. Perhaps most striking, however, was that prototype diagnosis consistently performed as well or better than both categorical *and* dimensional (number of criteria met) *DSM-IV* diagnoses in predicting external criteria widely viewed as central to the validity of a diagnostic system (see Robins & Guze, 1970), such as family history and adaptive

functioning. These single-item ratings were slightly superior to *DSM-IV* diagnosis in predicting efficacy of both psychotherapy and antidepressant medication. (In all cases, the empirically derived prototypes outperformed all of the alternative diagnostic systems.) In data analyses just completed (in preparation), clinicians' prototype ratings of single-sentence summaries of each of the 10 personality disorders taken from the text of the *DSM-IV* outperformed their own diagnostic judgments when adhering to *DSM-IV* decision rules (evaluating each symptom and counting the number of criteria met for each disorder).

Similar findings have been obtained with prototype ratings of mood, anxiety, and eating disorders, suggesting that the advantage of simple prototype ratings over complex symptom-counting algorithms are not specific to personality disorder diagnosis (Westen, Bradley, & Thompson-Brenner, 2004). Research just completed suggests that clinicians' prototype ratings may not only be valid predictors of external criteria but may also be made with substantial reliability in everyday practice. In a preliminary, small *N* study ($N = 37$), Westen, Bradley, and Hilsenroth (2005) asked advanced graduate student psychotherapists to make prototype ratings of personality disorder diagnoses derived empirically using Q-factor analysis (Westen & Shedler, 1999b) after their first 4 hours of clinical contact with the patient. A second advanced graduate student then watched the same 4 hours on videotape and made independent prototype ratings. Convergent validity correlations (e.g., histrionic ratings made by the therapist and the second rater) ranged from $r = .54$ to .89, with a median of .69, suggesting substantial agreement using a single-item measure applied to relatively unstructured (psychotherapy) data. Discriminant validity coefficients (e.g., the association between the therapist's ratings of histrionic personality features and the second observer's ratings of narcissistic features) hovered around zero. Intraclass correlation coefficients were in the same range, suggesting that clinicians could agree not only on the rank-ordering of patients on each prototype but on the absolute magnitude of resemblance to each prototype.

These data are clearly preliminary, and they do not tell us whether less-experienced clinicians could have done as well. However, taken together with the studies described above (and with other studies suggesting the possibility of reliable prototype ratings of this sort; e.g., Burisch, 1984; Elkin et al., 1995), they suggest that clinicians may be able to make reliable diagnostic judgments, and that these judgments are predictive of a range of variables supporting their validity across multiple samples and diagnostic classes.

Clinical Judgment and the Interpretation of Meaning

We turn now to a very different but central goal of clinical judgment, interpretive understanding. In a central tract on the philosophy of the social sciences, Max Weber (1949) distinguished what he called "adequacy at the level of meaning" (interpretative understanding of the subjective meaning of an action to a person or group) and "adequacy at the level of explanation" (what today we would call statistical prediction). He argued that a science devoted to the study of a meaning-making species such as our own must achieve both an understanding of the internal experience that leads people to do what they do, which inherently involves interpretation, and causal explanation, which is most convincing when it is probabilistic and predictive. From Weber's point of view, being able to predict the relation between two variables without understanding the subjectivity of the actor that mediates the relation is a useful step, just as being able to interpret the meaning of a psychological event without being able to make any probabilistic statements is useful, but each form of understanding is ultimately inadequate without the other.

In psychology, with our emphasis on significance testing and preferential tolerance for Type II over Type I errors, we have tended to judge theories and methods only on the

basis of Weber's second criterion. In many respects, this is sensible. An empirical discipline with quantifiable variables, effect sizes, and confidence intervals that can be estimated with some precision is an unwelcome host for the postmodern plague that has infected much of the humanities, cultural anthropology, and some circles in psychoanalysis. Disciplines for which hermeneutic exegesis of texts is a central component tend to be vulnerable to a proliferation of alternative exegeses, none of which can claim any truth value above the others, and hence to postmodern epistemological nihilism.

Although Meehl understood well the biases to which an interpretive mind is prone (and the indeterminacy of interpretive understanding), we suspect he would not have preferred a formula that could predict the number of first-person pronouns in Shakespeare's *King Lear* from the number in *Othello* and *Richard III* to the thoughtful analysis of a genuine expert on Shakespeare—unless of course, the latter was opining on the number of pronouns in *Lear*. Developments in the neurosciences suggest why Meehl never, in fact, jettisoned his belief in the value of interpretive understanding: Much of who we are and what we do resides in implicit cognitive–affective–motivational networks whose meaning is inaccessible to introspection and can only be accessed indirectly (Weinberger, in press; Westen, 1998).

Where networks are shared, researchers can observe their functioning using implicit tasks that minimize interpretive activity. For example, to the extent that *bird* and *robin* are associatively linked for most English speakers, we can reliably assess that linkage using priming tasks with reaction-time as a dependent variable. Where associations are idiosyncratic, however, as they are for everyone to some degree and particularly vis-à-vis psychopathology (which is why the behavior appears "abnormal"), we need to *identify* and *chart the topography* of networks that diverge from the ordinary—that is, to observe the idiosyncratic rather than the shared networks of thought, feeling, and motivation that regulate the person's mental life and behavior (see Westen, 1998; Westen, 1999; Westen, Feit, & Zittel, 1999). This, we suspect, is the reason for Meehl's tenacious belief (Paul E. Meehl, personal communication, May 2002) in the importance of eliciting patients' associations in therapeutic work. We can certainly establish group differences using such paradigms as emotional Stroop and lexical decision tasks (e.g., between depressed and nondepressed individuals, or between snake phobics and spider phobics; see, e.g., Teachman, Gregg, & Woody, 2001), but to understand *this particular depressed person*, the devil is in the details, and the details are often in associative links that take time and expertise to identify.

To what extent particular kinds of clinical training and experience foster the ability to "read" implicit associational networks (or, more broadly, to increase this aspect of "emotional intelligence") is an empirical question, but the fact that clinicians took the existence of such networks as axiomatic and tried to develop strategies for understanding and changing them a century before researchers did should give us pause before dismissing the interpretive activity of the clinician. Irrespective of the babble of (explicit) theories generated from clinical observation over the last century, from everything we know about implicit procedural learning, it is difficult to imagine that over years of clinical experience clinicians fail to encode *some* important regularities of their environment in implicit "grammars" of interpretation.

We will never be able to eradicate the fallibility (or artistry) inherent in interpretive understanding in psychotherapy, any more than we can eradicate it in musical performance. But we *can* improve the probability of valid interpretive activity in the clinical setting, in three ways. The first is to test and refine the theoretical frameworks that guide interpretative activity (e.g., testing hypotheses about how associations work; about how particular kinds of attachment experiences in childhood shape subsequent patterns of

thought, feeling, or interpersonal behavior; about how people regulate their emotions). As Kurt Lewin (1951, p. 169) said, there is nothing as practical as a good theory.

The second is to assess the validity of aggregated variables (e.g., diagnostic judgments) derived from a *pattern* of clinical inferences instead of assessing the validity of each specific inference. To put it another way, rather than assuming error-free clinical "items" (individual interpretations of ongoing clinical material), we can test the validity of the output of a series of clinical inferences aggregated to maximize clinical "reliability." For example, using a systematic clinical diagnostic interview designed to refine and formalize the kinds of interviewing practices used by experienced clinicians (Westen, 2003), experienced clinical observers can achieve remarkably high reliability in assessing narcissism ($r \geq .80$), even if they never ask patients if they are grandiose, entitled, readily enraged by feeling slighted, and so forth (Westen & Muderrisoglu, 2003, in press). The interview requires an experienced clinical interviewer to elicit narratives and observe the patient's behavior. What is striking about these findings is that clinicians can achieve high reliability across cases even when the questions they ask vary from patient to patient as long as they use a standard item set (e.g., a Q-Sort) to quantify and aggregate their inferences.

We do not believe, however, that clinicians only become capable of aggregating data at this level of inference if given a particular item set. Rather, we suspect that good clinicians practice a kind of "intuitive psychometrics" all the time: They discard inferences about the meaning of an idiosyncratic word choice, or about an unusual interpersonal interaction, if similar phenomena do not emerge with some regularity. In other words, they try to identify the central tendency in a sea of error and to assess internal consistency of their observations using what one might think of as an "intuitive alpha," which indexes the extent to which they can have confidence in inferred symptoms or personality attributes.[2] Of course, intuitive calculations of this sort can never be as accurate as the more precise calculations necessary for research. However, by quantifying clinicians' judgments using appropriate instruments, we can obtain more exact estimates of reliability and correlate these quantified observations with external criteria (e.g., indices of adaptive functioning, pooled informant reports) to assess the validity of a given clinician's judgments or a set of interpretive procedures.

This leads to the third way that we can test and improve the validity of clinical interpretive activity: by identifying those clinicians whose inferences, quantified reliably, show the strongest, most predictable pattern of correlates. We can then apply an expert systems approach to try to identify the procedures they are using, whether or not they themselves have explicit access to these procedures, and to teach other clinicians how to make similar inferences. We suspect that certain personality attributes are also predictive of the capacity for interpretive accuracy and hence useful in selecting potential clinicians, such as general (or perhaps verbal) intelligence, empathy, access to emotion, complexity of mental representations of people and relationships, and openness to experience.

Clinical Judgment and Therapeutic Intervention

In the most general sense, the practices evolved by clinicians over the last century "work": The effect size first estimated by Smith and Glass (1977) for general psychotherapy

[2] Clinicians may even not be aware of doing so. Work on implicit learning (e.g., Lewicki, 1986; Weinberger, in press) indicates that adults and children extract such central tendencies all the time in complex social situations. It would be surprising if practicing clinicians were inferior to 5-year-old children in this capacity. Indeed, as Goldberg (1991) has shown, one can often derive from clinicians' judgments regression equations that capture many of the rules they are using to make these judgments and then outperform the same clinicians by applying these equations without error to new cases.

effectiveness as compared with control conditions (roughly .85 in *SD* units) has stood the test of time and compares favorably to many other medical procedures (see Meyer et al., 2001; Wampold, 2001). This effect size is within the range (and generally in the middle of the confidence intervals) of effects obtained in randomized controlled trials (RCTs) for many disorders (see Westen, Novotny, & Thompson-Brenner, 2004).

The extent to which clinicians should rely on statistical algorithms rather than clinical judgment in selecting interventions with a given patient is the latest incarnation of the clinical–statistical debate and probably the most hotly contested version of that debate since the firestorm set off by Meehl's (in his words) "disturbing little book" written in 1954 (Meehl, 1954/1996). Advocates of the empirically supported therapies (EST) movement argue that we now have enough data to distinguish empirically supported from unsupported therapies for the most prevalent *DSM-IV* disorders (at least those coded on Axis I), and that clinicians should therefore limit their interventions to those on the EST list (Chambless & Ollendick, 2000). The EST movement has had a substantial impact on clinical training and practice, with many institutions now instructing doctoral students, interns, and postdoctoral fellows exclusively in techniques that have passed muster in controlled clinical trials or even in the implementation of the specific manuals used in these studies. Declarations of "treatment of choice" are common in the literature, with many suggesting that clinicians who do not follow these guidelines are providing poor care or perhaps even behaving unethically.

No one cognizant of the clinical–statistical debate (and of the cornucopia of fringe therapies available to an unwitting public) can doubt the importance of applying quantitative methods to treatment decisions in mental health (or any other health discipline). One can, however, question whether the EST movement represents the "statistical" side of the clinical–statistical debate as well as many of its advocates claim. Critics have raised a number of objections, the most important on empirical grounds. We will not attempt to adjudicate any of these issues here, but will simply present them in bulleted form:

- A positive therapeutic alliance and other common factors (mechanisms of therapeutic action common to virtually all therapies) tend to account for substantially more variance in outcome than specific treatment effects for most disorders studied (see Lambert & Bergin, 1994; Weinberger, 1995).
- Meta-analytic data consistently find that comparisons of two or more active treatments tend to produce modest effect sizes (absolute value around $d = .20$), particularly when both treatments are bona fide treatments (rather than pseudotreatments designed as sparring partners for "real" contenders; see Luborsky, Barton, & Luborsky, 1975; Luborsky et al., 2002; Wampold et al., 1997).
- Much of the variance in outcome widely attributed to specific treatment effects can be accounted for by allegiance effects, such that the treatment preferred by the investigator in a given study virtually always "wins" (Luborsky et al., 1999).[3]
- Most studies in the EST literature have applied extensive and idiosyncratic screening criteria that excluded many or most patients who presented with the symptom

[3] One might legitimately question whether some part of this variance accounted for by investigator allegiance can in fact, be attributed to investigators developing allegiances to treatments they have observed to work. However, empirically, one can list on the fingers of one hand the number of treatment researchers who have fundamentally changed their treatment orientation based on the results of their own clinical trials. Interestingly, we can think of many more *clinicians* who have fundamentally changed their views based on the data of *clinical* observation, such as Aaron T. Beck and Albert Ellis, who developed cognitive approaches to psychotherapy based on their clinical observation as psychoanalysts.

or syndrome under investigation, rendering generalizability uncertain (Westen & Morrison, 2001).

- Researchers have systematically avoided comparing experimental treatments with treatments widely used in the community (or treatments provided by clinicians identified empirically as those who obtain the best outcomes on average or clinicians nominated by their peers as experts), so that we do not know whether ESTs, which generally lead to improvement and recovery rates of 20 to 50 percent, are superior to outcomes in clinical practice (Westen, Novotny, et al., 2004, 2005).

From the point of view of the clinical–statistical debate, four points are of note here. First, given the lack of consensus in the literature, and the strong empirical arguments on both sides of the EST debate, the decision to *use* a formula in a given clinical instance to select an intervention is as "clinical" in Meehl's sense of the term (informal, subjective) as the decision to reject it. For example, EST researchers have come to the consensus that if a patient presents with depression, the appropriate psychosocial treatment is 16 sessions of cognitive or interpersonal therapy (Hollon, Thase, & Markowitz, 2002). Yet the average study of ESTs for depression has excluded over two thirds of the patients who presented for treatment, including those with suicidal ideation, substance abuse, and a range of other common features and comorbid conditions (Westen & Morrison, 2001). So how does a clinician decide whether this body of literature applies to this particular patient?

An EST advocate would suggest that, because some data are better than no data, clinicians should start with what has been validated and figure out what to do next if none of the extant ESTs works. Yet the available data suggest that the majority of carefully selected patients who undergo 16 sessions of cognitive or interpersonal therapy for depression (the treatment length prescribed in the manuals) administered by highly trained and supervised therapists in clinical trials fails to improve, remains symptomatic at termination, and relapses or seeks further treatment within 18 months (Westen & Morrison, 2001). In light of these dismal outcome statistics, and the fact that no one has ever compared these treatments with treatment in the community by expert practitioners, the assertion that clinicians should start with one of these manuals seems to us a "clinical" judgment with a low probability. It is unfortunate that researchers made the collective decision over the last 20 years to study only brief trials of only two treatments for a subset of poorly specified depressed patients and to avoid comparing them with treatments provided by experienced practitioners, because doing so has rendered the question of how to treat any given depressed patient in practice largely a matter of faith and opinion.

Second, advocates of ESTs often liken the choice of using versus not using an EST to statistical versus clinical prediction (e.g., Wilson, 1998), with the former relying on replicable research and the latter reflecting clinical opinion. However, one of the conditions Meehl (1954/1996) outlined in which statistical prediction will not outperform intuitive judgments by presumed experts is when a formula would be premature because of inadequate knowledge of relevant variables. To date, researchers have tested a very limited subset of possible interventions, namely those most congenial to brief treatments and to investigators whose theoretical meta-assumptions led them to believe (unlike Meehl) that psychopathology is highly malleable, distinct from personality, and hence likely to respond to brief treatments (Westen, Novotny, et al., 2004). Thus, a typical formula would include only three or four dichotomous variables (presence or absence of exposure, other techniques aimed at behavior change, cognitive restructuring, and interpersonal interventions), which only occasionally predict differences in treatment response.

An alternative strategy, truer to the spirit of Meehl's argument and the research literature on statistical prediction, would extend the range of scientific method to the

selection of interventions to test, measuring outcome and interventions strategies in large samples of patients treated in the community (as well as in the laboratory). Using an instrument with a sufficient number of intervention variables (e.g., the Psychotherapy Process Q-set; Ablon & Jones, 2002), one could then identify variables that predict positive outcomes for patients sharing a common problem (e.g., clinically significant depression) at clinically meaningful follow-up intervals. Researchers could use these findings to develop treatment protocols empirically and then subject them to experimental investigation. Using clinical practice as a natural laboratory in this way would transform the large variability of intervention strategies in the community from a source of realistic concern to a source of hypotheses and statistical knowledge linking interventions to outcome.

A third point worth noting regarding the applicability of the clinical–statistical debate to the EST debate pertains to the "clinical" nature of research decisions that have shaped psychotherapy research over the past two decades. Until the last decade, psychotherapy researchers considered the manipulation used in a given study to be one among a large number of possible ways to operationalize a principle or technique (e.g., exposure). As in most experiments, the particular manipulation chosen represented a sample of a broader population of possible operationalizations of a construct. In the EST era, however, manuals are viewed as *defining* treatments, not exemplifying them (see Westen, Novotny, et al., 2004). Clinicians are then exhorted to use the manual as written, not to draw inferences about general principles that might be applied in different ways to different patients or problems (see Beutler, 2000; Rosen & Davison, 2003). Often this exhortation includes the warning that clinicians should fight the impulse to choose from different manuals or improvise around a manual to fit a given patient, given that the clinical–statistical prediction literature has shown that such clinical judgments are generally unproductive.

In fact, however, researchers generally solidify the interventions embodied in treatment manuals long before testing them, using precisely the kinds of intuitive, informal processes Meehl termed *clinical*. That is, they use their best hunches to decide what interventions to include, how to operationalize them, how long treatment should last, and so forth, before ever conducting their first controlled trial. Only rarely do manuals undergo much significant change *after* their first successful clinical trial because no one wants to alter the "formula" that produced success, and the data often do not allow the researcher to identify which of many interventions that constitute the "package" (delivered in only one of multiple possible orders) were responsible for the obtained outcomes. Whether the empirically tested clinical hunches of a handful of like-minded researchers are better than the untested clinical hunches of several thousand experienced full-time clinicians is unknown. Personally, we would bet on reliability theory that the aggregate outcome of a thousand clinicians will outperform the aggregate outcome of treatments devised by a small number of researchers,[4] except where the researchers made creative use of basic science research unknown to most clinicians to generate a truly novel clinical innovation (e.g., Barlow's Panic Control Therapy; Barlow, 2002).[5] This is particularly the case given that researchers have generally imposed on themselves one of the same handicaps that severely limits clinical prediction, namely the lack of self-correcting feedback, in their failure to follow up their samples at intervals (e.g., 2 or 3 years) appropriate for disorders that are often slow to change or recurrent by nature.

---

[4] This is assuming, of course, that their errors are, if not randomly distributed, at least not correlated extremely highly. Given the multitude of theories guiding even clinicians who share a theoretical orientation, we suspect this assumption is reasonable enough.

[5] Fortunately, there is no need to bet; this is an empirical question that could be readily tested.

That the advantage of laboratory-generated treatments based on creative application of basic science may largely emerge only when researchers do something novel—that is, when they *do not* follow a formula—leads us to a paradox, because, as we have argued elsewhere (Westen & Weinberger, 2004), the advantage of statistical over informal, subjective decisions applies as much to experienced researchers as to experienced practitioners (who are both *clinicians* in Meehl's terms, i.e., purported experts). Meehl mercilessly attacked the foibles of clinicians, but we suspect he could have equally focused on the foibles of researchers. Indeed, if we were to apply the widely believed take-home message of the clinical–statistical debate (and the EST movement) to research, we would straightjacket researchers, requiring them to avoid creative application of scientific method in any given study in favor of manipulations or measures shown in prior research to be successful, given that any given hunch about why "this study is different" is vulnerable to all the biases involved when a clinician imagines that "this case is different." Unfortunately, there is no statistical algorithm for determining when to abandon a given algorithm or search for new ones. Had Einstein accepted formulas that had worked reasonably well for over a century, we would still be Newtonians. And we would argue, as suggested above, that the rote application of a set of methodological choices believed by researchers to have "worked" in psychotherapy research (and the enforcement of such choices through restriction of the kinds of studies that could obtain funding) has set back scientific knowledge about effective treatments for disorders such as depression by decades.

Conclusions: The Role of Clinical Judgment in Knowledge Generation

We have focused in this article on what clinicians may be able to do in everyday practice and have suggested that if we shift our focus from prediction at high levels of inference to (a) judgments at a moderate level of inference, (b) contexts in which clinicians are likely to develop expertise (diagnosis, interpretation of meaning, and intervention), and (c) conditions that optimize the expression of that expertise (e.g., use of psychometric instruments designed for expert observers), we may begin to see why Meehl never abandoned his belief in clinical expertise despite its vulnerabilities to all manner of mental shenanigans. We have suggested, in fact, that Meehl could not abandon the uncertainties of clinical judgment in the consulting room for the same reason researchers cannot abandon the uncertainties of "clinical" (i.e., expert, informal, subjective) judgment in the laboratory: Truth does not reveal itself without interpretation. The choice of what hypotheses to pursue, using what methods, is inherently a clinical decision, however informed (as it should be informed) by the available quantitative evidence. In this sense, the clinical–statistical debate is a particular incarnation of a central question of epistemology: how an imperfect mind can distinguish with any probability its own workings from the object of its investigations (see Meehl, 1983). Meehl's tentative solution, as an empiricist as well as an admirer of both Sigmund Freud and Albert Ellis, was to recommend that we stay abreast of the available data while remaining vigilant to both the emotional and cognitive biases to which human minds, including clinical minds, are prone.

We cannot conclude, however, without addressing one last context in which clinicians have laid claim to expertise: in generating psychological knowledge (see also Meehl, 1967,[6] on theory mediation). A century ago, Freud derived a theory of personality, psychopathology, and treatment largely from a clinical observation, and many approaches to psychotherapy today remain moored primarily in clinical judgment. Our goal in these

[6] This article was originally reprinted in Meehl, P.E. (1973). Psychodiagnosis: Selected papers (pp. 165–173). Minneapolis: University of Minnesota Press. The book itself was reprinted as Meehl, P.E. (1977). Psychodiagnosis: Selected papers. New York: Norton.

final pages is not to revive unbridled clinical theory-building, but to suggest several ways clinical observation may indeed have a place at the scientific table. We focus on six contexts in which clinicians may contribute to psychological knowledge: hypothesis generation, coherence testing, relevance testing, identification of disconfirming instances, item generation, and quantified observation. We cannot describe any of these in detail but will simply lay out some of the relevant issues.

The first way clinicians may contribute to psychological science is through hypothesis generation. It seems to us highly unlikely that people bright enough to get into highly competitive doctoral programs in clinical psychology learn nothing of consequence from their clinical training and develop nothing but self-sustaining biases from mucking around in the minds of other people for months or years at a time. Empirically, perhaps the best evidence is the fact that clinicians assumed the pervasiveness and significance of unconscious processes a century before researchers grudgingly admitted their existence; wrote about the complexity and importance of mental representations of the self and others in psychopathology decades before social cognition researchers discovered them (even applying the schema concept to self-representations; Sandler & Rosenblatt, 1962; Westen, 1992); and understood that attitudes can be complex, ambivalent (associated with multiple affects, and not just positive or negative valence), and primed outside of awareness long before researchers who spent their lives studying attitudes did (Westen, 1985).

No doubt clinicians of every theoretical persuasion engage their patients in what could be described from one point of view as shaping and from another as a *folie à deux*, in which clinicians selectively reinforce their patients for accepting the wisdom of their theoretical biases. This surely limits what both members of the therapeutic dyad learn from clinical hours, just as researchers' tendency to test only hypotheses they believe to be true at a subjective probability of, say, $p < .05$, limits what they learn, and it certainly sets severe constraints on the clinic as a venue for hypothesis *testing*. (Personally, we keep waiting for the good fortune of encountering one of those patients, widely presumed to be ubiquitous, who are so pliable, and whose pocketbooks so plentiful, that they believe everything we tell them and are willing to stick around for the duration of the loans on our BMWs to explore the vagaries of their Oedipus complex.) Vis-à-vis hypothesis generation, however, it is hard to imagine a better perch for observation than clinical practice (see Meehl, 1995b). And empirically, most important constructs in clinical psychology and psychiatry had their roots in clinical immersion (e.g., major depression, schizophrenia, anorexia nervosa, personality disorders, cognitive therapy).

The second way clinicians may contribute to psychological science is through what might be called *coherence testing*. The philosopher and cognitive scientist Paul Thagard (2000) has used the term "explanatory coherence" to describe the way people (including scientists) equilibrate to judgments or solutions to problems that maximize goodness of fit with the totality of cognitive constraints on what they might believe. The "hardest" constraints in science are hard data (e.g., the presence of certain findings in methodologically strong studies), but other constraints include less-definitive data (e.g., from case studies, or from correlational studies that do not definitively establish causation) and webs of related observations and hypotheses that make more or less sense in light of one or another potential solution to the problem at hand. Thagard has developed computational models that successfully simulate both everyday and scientific judgment using this approach, derived from connectionist models in cognitive science.

With respect to explanatory coherence, clinical observation, particularly when coupled with knowledge of relevant scientific literatures (e.g., basic science on emotion), has the strong advantage of providing a contextual "forest for the trees." For example, during the behaviorist era, most clinicians rejected out of hand the metatheoretical assumption

that thoughts and emotions are irrelevant or epiphenomenal because this principle had minimal explanatory coherence in the context of all they observed in the consulting room. Researchers, in contrast, could draw only on their everyday experiences to provide a broader context (or counterweight) for judging the viability of the hypotheses they were testing in the laboratory and correspondingly took 50 years to recognize the limitations of their meta-assumptions.

Or consider the cognitive revolution that followed the behavioral one in psychology. For 30 years, researchers pursued a serial-processing approach to memory and problem solving (i.e., a model positing that mental contents are processed one at a time in short-term memory) that largely ignored the role of emotion in both driving and distorting decision making. Although researchers have jettisoned pure serial processing models, and have expended considerable energy studying *cognitive* biases that can affect reasoning, the virtual absence of the word "affect" from the most recent reviews of the literatures on cognition, judgment, and decision making in the *Annual Review of Psychology* (e.g., Hastie, 2001) is a striking omission. Most clinicians would find most contemporary decision-making models uncompelling despite the strong empirical evidence supporting them, because such models would not allow them to generate reasonable hypotheses with explanatory coherence for the most basic clinical phenomena (e.g., how people choose whether to stay in relationships about which they are conflicted, how they recount and make sense of negative interpersonal encounters at work). Although one might (and should) shake one's finger in consternation at clinicians for not reading the *Annual Review of Psychology*, we would suggest, equally, that researchers who want to understand decision making might gain something from talking to people who study it in real time, in real life, 40 hours a week.

A third and related context in which clinicians might contribute to psychological knowledge is through what might be called *relevance testing*. We suspect that many treatment researchers could have gained substantially from talking with, rather than devaluing, clinicians over the last 20 years. Although part of the reason clinicians have been slow to embrace ESTs reflects the failure of many clinicians to attend to empirical research, equal blame can be laid at the doorstep of researchers, who never bothered to ask clinicians basic questions that might have been of use to them, such as what problems their patients actually bring to treatment, whether most patients meet criteria for a *DSM-IV* Axis I disorder, whether their patients tend to have multiple problems rather than a single disorder, how long it generally takes to treat various disorders successfully, and so forth. This could have been done quantitatively, by asking clinicians to describe, for example, their most recently terminated patient of a certain sort (e.g., a patient who presented with depression as an initial complaint), to minimize biases in sampling or reporting (see, e.g., Morrison, Bradley, & Westen, 2003; Thompson-Brenner & Westen, in press-a,b). The failure to do so has led to treatments that many clinicians consider of little relevance for most of their patients (see Westen, Novotny, et al., 2004).

A fourth role of clinicians in knowledge generation is the identification of disconfirming instances. As Hume noted years ago, all knowledge is probabilistic, and a single disconfirming case can provide extremely useful information for the refinement or disconfirmation of hypotheses. Hume noted how observation of a single black swan could shatter a compelling probabilistic hypothesis that all swans are white even after observing 100 consecutive white swans. As one of us observed in the heyday of the attributional reformulation of the helplessness theory of depression (Westen, 1985, 1991), depressed narcissists (e.g., middle-aged narcissists coming to grips with the reality that their grandiose dreams will never come true) are the "black swans" of learned helplessness theories. Treatment barely begins with narcissistic patients until they start to attribute

the causes of their misfortunes to their own stable traits, rather than blaming their difficulties on others.

A fifth context in which clinicians can contribute to knowledge generation is through item generation.[7] Just as clinical observation can be a wonderful vista from which to frame hypotheses, it can also be a useful vantage point for generating items for measures of clinically relevant constructs. A good example is the widely used Psychopathy Checklist-Revised (PCL-R; Hare, 2003). Not only the construct of psychopathy, but the items from the PCL-R were derived directly from the work of Cleckley (1941) based on his immersion in unstructured clinical observation of psychopaths. The same is true of many widely used instruments, such as the Beck Depression Inventory (Beck et al., 1961). By definition, the more one structures the context (as in an experiment) or the response options of respondents (as in a questionnaire), the less one can learn about things one did not already believe to be important or true (see Westen, 1988; Westen & Gabbard, 1999). Conversely, the less one structures the situation (as in immersion in clinical hours), the less one can conclude about causation or even patterns of covariation (as in the subsequent empirical finding, which could not have been readily known by Cleckley, that psychopathy is a multifactorial construct). This is precisely why clinical observation can be useful in the context of scientific discovery but much less useful in the context of scientific justification (hypothesis testing).[8]

A final way clinicians may contribute to psychological knowledge is as informants, through quantified observation. We will not belabor the point, which we have made extensively elsewhere (e.g., Westen & Shedler, 1999a), but in relying primarily on the self-reports of undergraduates, psychiatric patients, and other lay informants, as a field we may have ignored one of the most useful sources of information about patients' psychopathology—experienced clinicians who know them well. Ironically, part of the reluctance to quantify the observations of experienced clinical observers using psychometric procedures that have proven so useful in quantifying the observations of lay informants reflects a misunderstanding of the clinical–statistical debate, most importantly a confusion of the nature of the informant (clinician vs. lay self-report) with a method of aggregating data (subjective vs. statistical) (Westen & Weinberger, 2004). These two variables—informant and method of aggregation—constitute orthogonal axes, each of which has "clinical" at one pole; this has led, we believe, to considerable confusion. As we have tried to show, clinicians do appear to be able to make reliable and valid judgments at moderate levels of inference if given appropriate methods of quantifying their judgments.

The passing of one of the brightest lights ever to shine (and, if the reader will excuse a clinical prediction, ever likely to shine) on our profession seems an appropriate time to rethink the question of what clinicians should and should not be able to do. Our goal in this article was not to excuse clinicians for their failure to heed Meehl's warnings about the limits of clinical judgment, which they do at their own peril and that of their patients. Nor was it to pillory researchers for falling prey to many of the same biases (e.g., confirmation biases), and displaying much of the same hubris that is widely attributed to clinicians. We suspect that on the bell curves of foolishness and narcissism, clinicians and researchers probably show relatively similar means and dispersions (although this is

---

[7] Meehl himself believed in the importance of clinical immersion for item generation. When one of us (D.W.) once wrote him to ask for his comments on a draft of a clinician-report instrument designed to assess subtle forms of subpsychotic thought disorder, he responded, after reading the items, that the item content could only have been generated by a seasoned clinician. (We *believe* he meant this in a positive light, but that, of course, is a clinical judgment.)

[8] As Meehl put it, "The clear message of history is that the anecdotal method delivers both wheat and chaff, but it does not enable us to tell which is which" (1995b, p. 1019).

an empirical question). Rather, our goal has been to resurrect the less-studied pole of Meehl's conflict about research and practice. If, after 50 years of research, we have largely failed to identify manifestations of expertise in the only domain ever studied in which experience appears to lead to the accretion of little more than bias and error, the flaw may lie not only in the mental habits of clinical practitioners but in our own biases or lack of creativity as practitioners of scientific method.

References

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

American Psychiatric Association. (1968). Diagnostic and statistical manual of mental disorders (2nd ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association. (1980). Diagnostic and statistical manual of mental disorders (3rd ed.). Washington, DC: American Psychiatric Press.

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

Ablon, J.S., & Jones, E.E. (2002). Validity of controlled clinical trials of psychotherapy: Findings from the NIMH Treatment of Depression Collaborative Research Program. American Journal of Psychiatry, 159, 775–783.

Barlow, D. (2002). Anxiety and its disorders (2nd ed.). New York: Guilford Press.

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., Ward, C.H., et al. (1961). An inventory for measuring depression. Archives of General Psychiatry, 4: 561–571.

Beutler, L.E. (2000). David and Goliath: When empirical and clinical standards of practice meet. American Psychologist 55: 997–1007.

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

Bradley, R., Hilsenroth, M., & Westen, D. (2004). Validity of SWAP-200 personality diagnosis in an outpatient sample. Unpublished manuscript, Emory University.

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

Chambless, D., & Ollendick, T. (2000). Empirically supported psychological interventions: Controversies and evidence. Annual Review of Psychology, 52, 685–716.

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

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. Science, 243(4899), 1668–1674.

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.

Elkin, I., Gibbons, R.D., Shea, M., Sotsky, S.M., Watkins, J., Pilkonis, P., et al. (1995). Initial severity and differential treatment outcome in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. Journal of Consulting & Clinical Psychology, 63, 841–847.

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

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, MN: University of Minnesota Press.

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, & 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. (2003). Manual for the Hare Psychopathy Checklist–Revised. (2nd ed.). Toronto, ON: Multi-Health Systems.

Hastie, R. (2001). Problems for judgment and decision making. Annual Review of Psychology, 52, 653–683.

Hollon, S.D., Thase, M.E., & Markowitz, J.C. (2002). Treatment and prevention of depression. Psychological Science in the Public Interest, 3, 39–77.

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

Jampala, V., Sierles, F., & Taylor, M. (1988). The use of DSM-III-R in the United States: A case of not going by the book. Comprehensive Psychiatry, 29, 39–47.

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

Lambert, M., & Bergin, A. (1994). The effectiveness of psychotherapy. In A. Bergin & S. Garfield (Eds.), Handbook of psychotherapy and behavior change (4th ed., pp. 143–189). New York: Wiley.

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

Lewin, K. (1951) Field theory in social science; selected theoretical papers. D. Cartwright (Ed.). New York: Harper & Row.

Luborsky, L., Barton, S., & Luborsky, L. (1975). Comparative studies of psychotherapies: Is it true that "everyone has won and all must have prizes"? Archives of General Psychiatry, 32(8), 995–1008.

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.

Luborsky, L., Rosenthal, R., Diguer, L., Andrusyna, T.P., Berman, J.S., Levitt, J.T., et al. (2002). The dodo bird verdict is alive and well—Mostly. Clinical Psychology: Science & Practice, 9, 2–12.

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

Meehl, P.E. (1967). What can the clinician do well? In D.N. Jackson & S. Messick (Eds.), Problems in human assessment (pp. 594–599). New York: McGraw-Hill.

Meehl, P.E. (1973). Psychodiagnosis: Selected papers. New York: Norton.

Meehl, P.E. (1983). Subjectivity in psychoanalytic inference: The nagging persistence of Wilhelm Fliess's Achensee question. In J. Earman (Ed.), Minnesota studies in the philosophy of science: Vol. X, Testing scientific theories (pp. 349–411). Minneapolis: University of Minnesota Press.

Meehl, P.E. (1995a). "Bootstraps taxometrics: Solving the classification problem in psychopathology." American Psychologist, 50(4): 266–275.

Meehl, P.E. (1995b). "Is psychoanalysis one science, two sciences, or no science at all? A discourse among friendly antagonists": Comment. Journal of the American Psychoanalytic Association. 43, 1015–1023.

Meehl, P.E. (1996). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence (New Preface). Lanham, MD: Rowan & Littlefield/Jason Aronson. (Original work published 1954)

Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L., Dies, R.R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. American Psychologist, 56, 128–165.

Morrison, C., Bradley, R., & Westen, D. (2003). The external validity of efficacy trials for depression and anxiety: A naturalistic study. Psychology and Psychotherapy: Theory, Research and Practice, 76, 109–132.

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

Robins, E., & Guze, S. (1970). The establishment of diagnostic validity in psychiatric illness: its application to schizophrenia. American Journal of Psychiatry, 126, 983–987.

Rosen, G.M., & Davison, G.C. (2003). Psychology should list empirically supported principles of change (ESPs) and not credential trademarked therapies or other treatment packages. Behavior Modification, 27, 300–312.

Sandler, J., & Rosenblatt, B. (1962). The concept of the representational world. Psychoanalytic Study of the Child, 17, 128–145.

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.

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. Mahwah, NJ: Lawrence Erlbaum.

Smith, M., & Glass, G. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752–760.

Spitzer, R.L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria: Rationale and reliability. Archives of General Psychiatry, 35, 773–782.

Teachman, B.A., Gregg, A.P., & Woody, S.R. (2001). Implicit associations for fear-relevant stimuli among individuals with snake and spider fears. Journal of Abnormal Psychology, 2, 226–235.

Thagard, P. (2000). Probabilistic networks and explanatory coherence. Cognitive Science Quarterly, 1(1), 91–114.

Thompson-Brenner, H., & Westen, D. (in press). A naturalistic study of psychotherapy for bulimia nervosa, Part 1: Comorbidity and treatment outcome. Journal of Nervous and Mental Disease.

Thompson-Brenner, H., & Westen, D. (in press). A naturalistic study of psychotherapy for bulimia nervosa, Part 2: Therapeutic interventions and outcome in the community. Journal of Nervous and Mental Disease.

Wampold, B.E. (2001). The great psychotherapy debate: Models, methods, and findings. Mahwah, NJ: Erlbaum.

Wampold, B., Mondin, G., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empiricially, "all must have prizes." Psychological Bulletin, 112(3), 203–215.

Weber, M. (1949). The methodology of the social sciences. Glencoe, IL: Free Press.

Weinberger, J. (1995). Common factors aren't so common: The common factors dilemma. Clinical Psychology: Science and Practice, 2, 45–69.

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

Westen, D. (1985). Self and society: Narcissism, collectivism, and the development of morals. New York: Cambridge University Press.

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

Westen, D. (1991). Social cognition and object relations. Psychological Bulletin, 109, 429–455.

Westen, D. (1992). The cognitive self and the psychoanalytic self: Can we put our selves together? Psychological Inquiry, 3(1), 1–13.

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

Westen, D. (1999). Psychodynamic theory and technique in relation to research on cognition and emotion: Mutual implications. In T. Dalgleish & M. Power (Eds.), Handbook of cognition and emotion (pp. 727–746). New York: Wiley.

Westen, D. (2003). Clinical diagnostic interview. Unpublished manuscript, Emory University, Atlanta, GA. Retrieved from www.psychsystems.net/lab

Westen, D., & Bradley, R. (2005). Prototype diagnosis of personality. In S. Strack (Ed.), Handbook of personology and psychopathology (pp. 238–256). New York: Wiley.

Westen, D., Bradley, R., & Hilsenroth, M. (2005). Reliability of prototype diagnosis of personality in clinical practice. Unpublished manuscript, Emory University.

Westen, D., Bradley, R., & Thompson-Brenner, H. (2004). A prototype approach to the diagnosis of mood, anxiety and eating disorders. Unpublished manuscript, Emory University, Atlanta, GA. .

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., & Gabbard, G.O. (1999). Psychoanalytic approaches to personality. In L. Pervin & O. John (Eds.), Handbook of personality: Theory and research (2nd ed., pp. 57–101). New York: Guilford.

Westen, D., Heim, A.K., Morrison, K., Patterson, M., & Campbell, L. (2002). Simplifying diagnosis using a prototype-matching approach: Implications for the next edition of the DSM. In L.E. Beutler & M.L. Malik (Eds.), Rethinking The DSM: A Psychological Perspective (pp. 221–250). Washington, DC: American Psychological Association.

Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies. Journal of Consulting and Clinical Psychology, 69, 875–899.

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. (in press). Clinical assessment of pathological personality traits. American Journal of Psychiatry.

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., Novotny, C.M., & Thompson-Brenner, H. (2005). EBP ≠ EST: Reply to Crits-Christoph, Wilson, and Hollon (2005) and Weisz, Weersing, and Henggeler (2005). Psychological Bulletin. 13, 427–433.

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. (2000). A prototype matching approach to diagnosing personality disorders toward DSM-V. Journal of Personality Disorders, 14, 109–126.

Westen, D., Shedler, J., & Bradley, R. (2004). A prototype matching alternative for diagnosing personality disorders in clinical practice. Unpublished manuscript, Emory University, Atlanta, GA.

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. (2004). When clinical description becomes statistical prediction. American Psychologist, 59, 595–613.

Widiger, T.A., & Clark, L.A. (2000). Toward DSM-V and the classification of psychopathology. Psychological Bulletin, 126, 946–963.

Wilson, G. (1998). Manual-based treatment and clinical practice. Clinical Psychology: Science & Practice, 5, 363–375.