Citation:
Meyer, G. J. (2002). Implications of information-gathering methods for a refined taxonomy of psychopathology. In L. E. Beutler & M. Malik (Eds.). *Rethinking the DSM: Psychological perspectives* (pp. 69-105). Washington, DC: American Psychological Association.

3

IMPLICATIONS OF INFORMATION-GATHERING METHODS FOR A REFINED TAXONOMY OF PSYCHOPATHOLOGY

GREGORY J. MEYER

Two things are required before clinicians can assess psychopathology or diagnose psychological disorders. First, one must have a framework to organize the relevant constructs that indicate pathology or disorder. These constructs may be clusters of subjectively experienced symptoms, behavioral acts, dichotomous taxonomies, continuous dimensions derived from factor analyses, or one of many other possible templates for classifying human functioning. Regardless of the framework, the constructs of the classification scheme provide a road map to the phenomena that a clinician should examine to identify health and illness.

A large portion of this book is devoted to identifying limitations with the dichotomous classifications that are applied to symptomatology in the current *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*; American Psychiatric Association, 1994). Another large portion of this book is devoted to explicating alternative, empirically defined, and empirically defensible frameworks that could replace or supplement the *DSM*.

This chapter does neither; I do not advance or address any particular framework of pathology. Instead, I focus on the second element that is required for a classification scheme, namely, the mechanisms or methods that allow one to gather relevant information about a patient (e.g., scales, interviews, performance tasks) to determine whether he or she has the characteristics that fall within a given framework of psychopathology. Thus, I focus on how one classifies or designates patients as having certain characteristics, not on what characteristics get classified or designated.

To ground the issues, I first draw from a large body of evidence that addresses the distinctions among sources of information gathering and the incompleteness of any single source. Next, additional data are used to exemplify how sources of information gathering have a profound impact on traditional *DSM* diagnoses. These data serve three purposes. First, from a psychometric perspective, they demonstrate how reliability and validity coefficients are on a single continuum that measures what can be considered "source overlap" or "criterion contamination." Second, the data demonstrate how the methodological factors associated with individual studies cause the research literature to generate wildly different estimates concerning the validity of the diagnostic information obtained from a given source. Third, and most important, the same data clearly demonstrate that diagnostic decisions can be highly source dependent and thus unstable and less than optimally valid representations of reality.

Finally, regardless of the specific content that defines pathology within any system developed to replace or supplement the *DSM*, I conclude that an empirically guided alternative should formally and systematically require clinicians who use the system to consider multiple, maximally independent sources of information prior to making diagnostic determinations. I review several strategies for doing so and in the process highlight the emerging evidence that indicates that these strategies improve the quality of information used in clinical decision making.

THE IMPORTANCE OF "HOW" RATHER THAN "WHAT"

One may wonder why the manner of assessing psychopathology is important. After all, most clinical training in psychology, psychiatry, and other mental health professions is oriented toward teaching the content to be considered for diagnostic or treatment purposes. As long as the content is clearly specified, shouldn't it be an easy matter to interview a patient and determine whether that particular phenomenological quality is present or absent? Psychologists, more than other mental health professionals, should recognize that the answer to this question is negative. It is not easy to obtain clear or definitive answers about psychological characteristics, and it never has been. The historical and contemporary evidence unambiguously indicates that it is virtually impossible to assess any psychological characteristic in a clear or definitive way.

As a discipline, we were first alerted to this conundrum more than 40 years ago (Campbell & Fiske, 1959). Since that time, important works have appeared with some regularity to remind us of the problem (e.g., Achenbach, McConaughy, & Howell, 1987; Cook & Campbell, 1979; Kagan, 1988; Michell, 1997; Mischel, 1968/1996; Nisbett & Wilson, 1977). Although these articles and books address many distinct points, they share a common thread, namely, that psychological content cannot be separated from the mechanisms or methods used to assess that content. With virtually all psychological phenomena, when the "same" construct or characteristic is assessed by distinct methods, different conclusions are obtained. For instance, to know the extent of depression in a girl seeking treatment, one would arrive at different conclusions depending on whether the source of information about her depression came from the child herself, her mother, her father, her teacher, her peers, or her prior therapist (see Achenbach et al., 1987).

For some reason, research and applied practice have been slow to embrace these findings, particularly with adults, where it is standard to obtain information from just the patient. Perhaps this is because the results are humbling. It would be comforting to think that conducting a skilled interview with a reasonably open patient would allow us to learn all that is necessary to reach a firm understanding and draw sensible conclusions about the patient. However, the data indicate that this is not so. Although impressions from interviews are certainly valuable, like every source of information, they are inherently limited and incomplete.

It is unlikely that anyone would strongly dispute the notion that impressions from interviews are limited and incomplete. After all, it would be preposterous to assert that they were perfect sources of information, just as it would be preposterous to assert that any source of information was perfect. What becomes important then is evidence on the extent to which different sources of information are incomplete. We should be very clear about the degree of association that exists between one source of information and another, particularly when considering the impact that diagnostic determinations in applied practice may have on the life of an individual. Knowing the degree of association between one source of clinical data and another should allow us to reach level-headed conclusions about whether our reliance on imperfect data sources really constitutes a salient problem in applied clinical practice.

RELIANCE ON PATIENT SELF-REPORT IN CURRENT DIAGNOSTIC PRACTICES

Before considering data pertaining to the relationship between different data sources, it is instructive to briefly review some issues related to diagnostic decision making. *DSM* diagnoses are fundamentally made by

collecting data from clinical interviews with the patient (or, in the case of a child, with the child's caregiver). Thus, even though *DSM–IV* mentions "associated laboratory findings" or "associated physical examination findings" that may help contribute to diagnostic determinations (e.g., memory testing or neuroimaging for dementia disorders, toxicology for substance use disorders, hypnotizability for dissociative disorders, polysomnography for sleep disorders), clinicians are rarely encouraged to obtain diagnostically relevant information from sources other than the patient. There are several exceptions. For instance, when assessing delirium disorders, *DSM–IV* states that patients may be incoherent and that "under these circumstances" (American Psychiatric Association, 1994, p. 125) it may be helpful to obtain information from family members or other informants. Patients with anorexia nervosa are said to frequently lack insight into their eating problems. Consequently, it is "often necessary" (p. 540) to obtain information from third parties to evaluate weight loss and other symptoms. For personality disorders, *DSM–IV* states that it "may be helpful" to obtain "supplementary information from other informants" (p. 630) to overcome the fact that patients with a personality disorder may not consider their personal characteristics to be problematic. Perhaps the strongest instructions for considering additional sources of information besides the patient's self-report come in the section on substance-related disorders. Although it is not formalized in any of the diagnostic criteria for these disorders, clinicians are told that they should consult additional sources of information "whenever possible" (p. 185) while they are obtaining a detailed substance use history. Despite these exceptions, but consistent with the interview schedules that have been developed over the past 20 years, making a diagnosis fundamentally revolves around obtaining information directly from the patient.

Historically, diagnoses were formulated from free-flowing clinical interviews that did not have a proscribed structure or a standard set of probe and follow-up questions. During an interview, clinicians tracked the patient's discourse and intermittently intervened by asking impromptu questions that followed up on hunches, or they solicited examples here and there to help gain a more complete and fleshed-out view of the patient's complaints. Unfortunately, the absence of a general structure to the interviews seemed to contribute to rather poor reliability. Different clinicians would often disagree about the same patient's core problems. For instance, across seven data sets obtained between 1956 and 1978, the average agreement on the *DSM* diagnoses from unstructured joint or separate interviews was $\kappa = .64$ ($N = 1{,}141$; 3–26 disorders considered; data were combined from Spitzer, Forman, & Nee, 1979; and Spitzer & Fleiss, 1974; although Study 5 was excluded from Spitzer & Fleiss because it compared clinical diagnoses to structured research diagnoses).

To deal with this problem, semistructured and fully structured interviews have been developed. Fully structured interviews provide specific probe and follow-up questions to address almost every contingency that may emerge in the interview. These schedules have gone so far to remove clinical judgment from the diagnostic interview process that they have been fully automated. Thus, in a demonstration of oxymoronic labeling, one can use "computer-administered diagnostic interviews" to make diagnoses. With the most highly structured interviews, different clinicians attending the same interview can produce almost exactly identical diagnoses. For instance, Wittchen (1994) reviewed data on the Composite International Diagnostic Interview (CIDI), a fully structured interview that can be administered by clinicians or by computer. Wittchen demonstrated that two clinicians attending the same interview produced near-identical results across 21 diagnostic categories, with a mean kappa of .92 ($N = 575$).

Semistructured interviews have also produced positive results when clinicians observe the same interview. For instance, across 3–22 Axis I disorders, semistructured joint interviews have produced an average diagnostic reliability of $\kappa = .72$ ($N = 514$; data were combined from Keller et al., 1995; Riskind, Beck, Berchick, Brown, & Steer, 1987; Segal, Hersen, & Van Hasselt, 1994 [using data from Table 1 that were otherwise unavailable]; Skre, Onstad, Torgersen, & Kringlen, 1991; and Steinberg, Rounsaville, & Cicchetti, 1990). Similarly, across 4–12 Axis II disorders, semistructured joint interviews have produced an average diagnostic reliability of $\kappa = .76$ ($N = 1{,}120$; data were combined from Bernstein et al., 1997; Gómez-Beneyto et al., 1994; Maffei et al., 1997; and Zimmerman, 1994, and the Zimmerman data were supplemented by published results in Loranger et al., 1994).

Although interclinician diagnostic agreement may well have improved with the advent of structured and semistructured interview schedules,[1] this shift brings about other important changes as well. As one moves from unstructured clinical interviews to semistructured interviews to fully

[1]The reliability values presented here can raise legitimate questions about whether more highly structured interviews improve reliability. If unstructured interviews produce an average kappa of .64, whereas semistructured interviews produce average values of .72 or .76, the change does not seem very dramatic—particularly when recognizing that all of the semistructured coefficients emerge from joint rather than independent interviews. Indeed, when independent interviews are conducted over short retest intervals, reliability declines for structured and semistructured interviews. For instance, Wittchen's (1994) review of the fully structured CIDI revealed an average kappa of .68 over a 1–3 day period (mean $N = 117$ over 25 disorders). For semistructured interviews, Williams et al. (1992) found an average kappa of .60 for Axis I disorders over a 1–14 day period ($N = 592$; 9–19 disorders) and Zimmerman's (1994; Table 2) review indicated an average kappa of .52 over the same time period for Axis II disorders ($N = 358$; 6–11 disorders). Although these findings paradoxically suggest that unstructured diagnostic interviews may be about as reliable as structured or semistructured interviews, the data reported here are imprecise and incomplete. For instance, a number of the samples that examined unstructured interviews had confounds that would artificially increase reliability values (e.g., cases of disagreement were not always included in Spitzer et al.'s, 1979, analysis), and none of these data emerge from an exhaustive search of the relevant literature. This limited survey is provocative, but a more definitive review is necessary to adequately address the issue.

structured interviews, there is a parallel shift in the emphasis given to the source of information used in the diagnostic process. Progressing along this continuum produces a decreasing reliance on the clinician's inferences and an increasing reliance on the patient's self-report.

Diagnostic determinations that are made from an unstructured clinical interview are certainly informed by the patient's self-report. However, the clinician is in the central role of synthesizing information and making inferences. As such, unstructured clinical interviews rely heavily on the skill and sophistication of the clinician. In contrast, clinical inference is completely absent from fully structured interviews. Here, patients are asked specific questions, and they provide specific answers. Endorsement of the presence or absence of a particular symptom is in the hands of the patient, not the clinician. This should be no surprise. Fully structured interviews were developed to remove clinical skill from the equation (see L. N. Robins, Helzer, Croughan, & Ratcliff, 1981).

As the reliability data presented above attest, fully structured joint interviews maximize interrater reliability. Regardless of the clinician conducting the interview or processing the information, virtually identical diagnoses are derived. This seems like it should be a good thing. Perhaps so. However, it also means that diagnostic determinations are much more source dependent because they rely exclusively on the perspective of the patient. Of course, patient perceptions can be marred by memory problems, confusion, a distorted self-image, the denial of or failure to recognize personal characteristics, and deliberate efforts to present oneself in a slanted and inaccurate manner (e.g., John & Robins, 1994; Malgady, Rogler, & Tryon, 1992; Meyer, 1997; Rogler, Malgady, & Tryon, 1992). Although most clinicians find structured or semistructured interviews to be helpful when evaluating Axis I symptomatology, clinicians almost across the board revert to an unstructured interview format when attempting to understand the broader matrix of the patient's personality (Westen, 1997). In essence, when the increased reliability offered by highly structured interviews is purchased, it is paid for with dependence on the patient as the primary source of diagnostic information. This is not necessarily a sound investment.

DATA ON THE CORRESPONDENCE BETWEEN DIFFERENT SOURCES OF INFORMATION

Meyer et al. (2001) organized a large array of data that examined the extent of association between distinct ways of assessing the same construct. The methods included structured interviews, semistructured interviews, unstructured interviews, self-reports, parent reports, spouse or significant other reports, peer reports, teacher reports, observed behavior, performance tasks, and inferences synthesized from multiple sources of information.

Table 3.1 presents those findings. Although the results do not emerge from an exhaustive search of all the relevant literature, they appear representative and generalizable, because they draw on data from more than 800 samples and more than 185,000 participants. When constructing this table, studies were excluded if they used aggregation strategies to maximize associations (e.g., self-reports correlated with a composite of spouse and peer reports; see Cheek, 1982; Epstein, 1983; Tsujimoto, Hamilton, & Berger, 1990), and moderators of agreement that may have been identified in the literature were ignored. Studies were also excluded if data sources were not reasonably independent. For instance, studies were omitted if patients completed a written self-report instrument that was then correlated with the results from a structured interview that asked the same or comparable questions in an oral format (e.g., Richter, Werner, Heerlein, Kraus, & Sauer, 1998).

A careful review of Table 3.1 indicates that distinct information emerges when distinct sources are consulted about the same psychological characteristic. This is evident from the relatively low to moderate associations that are seen between independent methods of assessing similar constructs. The findings hold for both children and adults, and they hold when various types of knowledgeable informants (e.g., self, clinician, parent, peer) are compared to one another or to observed behaviors and task performance.

For instance, the information obtained from children and adolescents has only moderate correspondence with the same information obtained from parents (table entries 1–4), teachers (table entries 8–10), clinicians (table entries 5 and 6), or other observers (table entry 7). Furthermore, the information obtained from parents, teachers, clinicians, or observers is largely independent; each of these sources has only moderate associations with the other sources (table entries 12–18 and 20–21). For adults, the information obtained from self-reports of personality and mood has only small to moderate associations with the same information when it is obtained from those who are close to the target person (table entries 23–25 and 29–30), from peers (table entries 26–28), from clinicians (table entries 31–34), from performance tasks (table entries 37–43), or from observed behavior (table entries 44–46).

The substantial independence between sources of information is not a phenomenon that is limited to rarified psychological research on personality; the independence pervades all aspects of human functioning and, as such, it clearly extends into the office of every clinician who attempts to treat or understand the psychopathology of his or her patients. Not only do patients, clinicians, parents, and observers have different views about psychotherapeutic progress or functioning in treatment (see table entries

TABLE 3.1
A Sample of Cross-Method Convergent Associations Indicating the General Accuracy of Single, Independent Sources of Information

Sources of Data and Constructs	*r*	κ	*N*
Children and Adolescents			
1. Self vs. parent: Behavioral and Emotional Problems	.29		14,102
2. Self vs. parent: Behavioral and Emotional Problems (Q-Correlations of profile similarity)	.29		1,829
3. Self vs. parent: Symptom Change in treatment	.19		199
4. Self vs. parent: DSM Axis I disorder		.24	1,136
5. Self vs. clinician: Behavioral and Emotional Problems	.14		1,079
6. Self vs. clinician: DSM Axis I disorder		.23	998
7. Self vs. clinical observer: Change in treatment	.28		199
8. Self vs. teacher: Behavioral and Emotional Problems	.21		9,814
9. Self vs. teacher: Behavioral and Emotional Problems (Q-Correlations of profile similarity)	.17		1,222
10. Self vs. teacher: Test Anxiety	.23		3,099
11. Self vs. aggregated peer ratings: Behavioral and Emotional Problems[b]	.26		8,821
12. Parent vs. teacher: Summed Behavioral and Emotional Problems	.29		29,163
13. Parent vs. teacher: Specific Behavioral and Emotional Problems	.16		1,161
14. Parent vs. teacher: Behavioral and Emotional Problems (Q-correlations of profile similarity)	.22		2,274
15. Parent vs. teacher: *DSM* Axis I disorder		.13	1,229
16. Parent vs. clinician: Behavioral and Emotional Problems	.34		1,725
17. Parent vs. clinician: *DSM* Axis I disorder		.39[a]	786
18. Parent vs. direct observer of child behavior: Behavioral and Emotional Problems	.27		279
19. Parent vs. cognitive test: Attention Problems	.03		451
20. Teacher vs. clinician: Behavioral and Emotional Problems	.34		1,325
21. Teacher vs. direct observer of child behavior: Behavioral and Emotional Problems	.42		732
22. Teacher vs. cognitive test: Attention Problems	.10		483
Adults			
23. Self vs. spouse–partner: Personality and Mood	.29		2,011
24. Self vs. spouse–partner: "Big Five" Personality Traits (Domains and Facets)	.44		1,774
25. Self vs. parent: Personality Characteristics (including the "Big Five")	.33		828
26. Self vs. peer: Personality and Mood	.27		2,119
27. Self vs. peer: "Big Five" Personality Traits (domains and facets)	.31		1,967
28. Self vs. peer: Job Performance	.19		6,359
29. Self vs. significant other: Attention Problems and Impulsivity	.22		202
30. Self vs. significant other: *DSM* Axis II personality disorder diagnosis		.12	768

Table Continues

TABLE 3.1
(*Continued*)

Sources of Data and Constructs	*r*	κ	*N*
31. Self vs. clinician: Treatment Related Functioning, Symptomatology, and Outcome	.29		7,903
32. Self vs. clinician: *DSM* Axis II personality disorder characteristics	.33[a]		2,778
33. Self vs. clinician: *DSM* Axis II personality disorder diagnosis		.18[a]	2,859
34. Self vs. clinician: *DSM* Axis I disorders		.34[a]	5,990
35. Self vs. clinician: "Big Five" Personality Traits (domains only)	.32		132
36. Self vs. supervisor: Job Performance	.22		10,359
37. Self vs. subordinate: Job Performance	.14		5,925
38. Self vs. cognitive test or grades: General Intelligence	.24		904
39. Self vs. cognitive test or grades: Scholastic Ability[c]	.38		8,745
40. Self vs. cognitive test: Memory Problems	.13		5,717
41. Self vs. cognitive test: Attention Problems	.06		522
42. Self vs. Thematic Apperception Test: Achievement Motivation[c]	.09		2,785
43. Self vs. Thematic Apperception Test: Problem Solving	.13		199
44. Self vs. Rorschach: Emotional Distress, Psychosis, and Interpersonal Wariness	.04		689
45. Self vs. observed behavior: Personality Characteristics	.16		274
46. Self vs. observed behavior: Attitudes	.32		15,624
47. Peers vs. observed behavior: Personality Characteristics	.15		264
48. Clinician vs. consensus best estimate: *DSM* Axis II Personality Disorder Diagnosis		.28	218
49. Significant other vs. significant other: Target Patient's *DSM* Personality Disorder Diagnosis		.32	386
50. Significant other vs. clinician: Target Patient's Depressive Signs and Symptoms		.13	141
51. Judgments from one source of test data vs. another: Personality, Needs, and IQ[d]	.12		158
52. Supervisor vs. peers: Job Performance	.34		7,101
53. Supervisor vs. subordinate: Job Performance	.22		4,815
54. Peers vs. subordinate: Job Performance	.22		3,938
55. Objective criteria vs. managerial ratings: Job Success	.32		8,341

Note: Citations and specific notes for each entry are provided in Meyer et al. (2001). *r* = Pearson correlation; κ = kappa coefficient; *N* = number of participants; *DSM* = *Diagnostic and Statistical Manual of Mental Disorders*; IQ = intelligence quotient. From "Psychological Testing and Psychological Assessment: A Review of Evidence and Issues," by G. J. Meyer, S. E. Finn, L. D. Eyde, G. G. Kay, K. L. Moreland, R. R. Dies, E. J. Eisman, T. W. Kubiszyn, and G. M. Reed, 2001, *American Psychologist, 56*, pp. 146–149. Copyright 2001 by the American Psychological Association.

[a]These coefficients are inflated by criterion contamination. For instance, in an effort to maximize cross-observer correspondence, one study went so far as to exclude the inferences that clinicians developed from their direct observations of the patient as a way to increase diagnostic agreement between patients and clinicians. [b]Because much of these data reflect the correlation between aggregated peer ratings and self-ratings, the coefficient is larger than would be obtained between self-ratings and the ratings of a single peer. [c]Result combines some data from children and adolescents with adults. [d]These studies were from the late 1950s and early 1960s. It is unclear whether the data may be different using more contemporary scoring and interpretive practices.

3, 7, and 31), but diagnoses have only moderate associations when they are derived from self-reported information or the information reported by parents, significant others, and clinicians (see table entries 4, 6, 15, 17, 30, 33, 34, 47, and 48). For instance, personality disorder diagnoses derived from self-report bear little resemblance to those derived from clinicians (κ = .18, N = 2,859; table entries 33) or to those based on semistructured interviews with significant others in the patient's life (κ = .12, N = 768; table entry 30).

DATA ON THE RELIABILITY–VALIDITY CONTINUUM AND THE POTENTIALLY EPHEMERAL NATURE OF DIAGNOSES

Given the purpose of this book, the findings on diagnostic (dis)-agreement across distinct sources of information deserve closer scrutiny. In this section, I review existing research on the accuracy or validity of diagnostic decisions derived from patients as the sole source of information. These data illustrate three points. First, although not central to the goals of this chapter, the data demonstrate a psychometric principle: Reliability and validity coefficients are on a single continuum that measures what can be considered "source overlap" or "criterion contamination." Second, the data show how the methodological decisions that are made when designing individual studies force the research literature to produce wildly different estimates concerning the accuracy of diagnoses obtained from a given source of information. Third, and most important, the data show how final diagnostic decisions can be incredibly source dependent. As a result of this source dependence, clinical diagnoses are often unstable and inaccurate representations of reality.

My main thesis here is that different sources of information provide distinct clinical data that are relevant to the diagnostic process. A hypothesis that flows from this is that as methods of information gathering become more distinct, the extent of association between different data sources will decline. Stated differently, to the extent that sources of information overlap or become confounded with each other, cross-source diagnostic agreement will become elevated. This can lead to an illusion of accuracy whereby a given source of diagnostic information is viewed as more accurate or trustworthy than is warranted.

To examine this illusion of accuracy in a diagnostic context, evidence about the accuracy of fully structured interviews is reviewed. Over the past 20 years, a fairly extensive literature has developed to determine the accuracy or validity of these measures. The two most prominent tools are the CIDI, which was mentioned earlier, and its direct predecessor, the Diagnostic Interview Schedule (DIS; L. N. Robins et al., 1981). The sole purpose of both instruments is to make diagnoses according to the *DSM* or the *International Classification of Diseases* (*ICD*). Because both instruments are fully structured, they are ultimately self-report scales (as evidenced by their ability to be administered by either an interviewer or a computer).

According to the hypothesis outlined above, one should expect that the DIS or CIDI produces higher validity coefficients with other methods of making a diagnosis when the alternative methods largely draw on the same source of information as the DIS and CIDI. Thus, when both the CIDI or DIS diagnoses and the external criterion diagnoses are largely derived from the patient's self-reported information, the CIDI or DIS diagnoses should display the highest levels of agreement with the external criterion diagnoses. This should occur, for example, when the clinician who makes the external diagnosis has also administered the CIDI or DIS and has not considered much information beyond that obtained during the CIDI or DIS interview. In contrast, CIDI or DIS diagnoses should display the lowest levels of agreement with external criterion diagnoses when the latter are derived from the most independent sources of data. This should occur, for instance, when criterion diagnoses are assigned by the consensus of clinicians who have observed hospitalized patients over time, interviewed them on multiple occasions (particularly with both structured and unstructured interviews), and solicited information from additional people who know the patient well (e.g., family members, therapists, friends).

The two research designs mentioned above fall at opposite ends of a methodological continuum of validity. This continuum indicates the extent to which criterion diagnostic decisions are confounded with the predictor diagnostic determinations obtained from the CIDI or DIS. In the language of psychometrics, this continuum addresses *criterion contamination*, which is present to the extent that a criterion is influenced or determined by the very information that is also used to predict it.

Most often, criterion contamination is identified when the results from a traditional test scale inform the criterion judgments that are used to validate the scale. For instance, if intelligence test scores are used to predict teacher ratings of intelligence, but the teacher ratings are completed after the teachers are exposed to the results of the intelligence test, the study would have criterion contamination and would produce artificially high evidence of validity for the intelligence test.

However, this is just one phenotypic manifestation of the underlying problem; criterion contamination emerges in other ways, too. What I focus on here is how artificially inflated estimates of validity or diagnostic accuracy can emerge as a function of the source of information that informs both the predictor and the criterion. This can be considered a source overlap problem. In terms of diagnoses, when criterion diagnoses are derived from the same source of information that generates the predictor diagnoses, the predictor and criterion information is confounded and should produce artificially high estimates of the predictor's validity or accuracy.

For instance, when the CIDI is administered to a patient by a lay interviewer and then shortly thereafter it is administered by a psychologist or psychiatrist, the predictor (i.e., diagnoses from the CIDI by the lay interviewer) and the criterion (i.e., diagnoses from the CIDI by the professional interviewer) are quite confounded because both sets of diagnoses draw on the same source of information for making diagnostic decisions (i.e., the patient's self-report). To the extent that both the predictor and the criterion drink exclusively from the same well, they are both similarly influenced by whatever was in that water. In the case of the fully structured CIDI, which is completely dependent on the patient's self-report, both the predictor diagnoses and the criterion diagnoses would be entirely determined by the same single source of information.

These ideas lead to expectations for what should be observed as one looks across the research literature. Those studies that are designed in such a way that the predictor and criterion rely most heavily on the same source of information should produce the largest estimates of validity or accuracy. To organize the literature examining CIDI or DIS diagnoses as the predictor relative to criterion diagnoses derived by some additional or alternative method, four categories of research methodology are delineated. These four methods vary in their degree of source overlap or criterion contamination and thus are expected to produce quite different estimates of CIDI or DIS validity.

First, high coefficients of agreement should be found when clinicians make the criterion diagnoses after having just attended or performed the CID or DIS interviews but supplement that information with the results from some additional questions designed to clarify remaining ambiguities related to the diagnostic criteria. If one thinks in terms of Venn diagrams, the criterion diagnoses would be drawn from a universe of information that encompasses and is just slightly larger than the CIDI or DIS universe of information (e.g., the outline of a dime inside the outline of a penny). In terms of relative proportions, the criterion diagnoses are almost completely determined by the same information that generated the CIDI or DIS diagnoses.

Second, moderately high coefficients of diagnostic agreement would be expected when the criterion diagnoses are derived from a highly structured interview that is administered on the same day as the CIDI or DIS but in the context of an independent interview. Using a highly structured interview for the criterion diagnoses ensures that the patient is the key source of information for both the predictor and the criterion. In addition, by administering the CIDI or DIS and the criterion interviews on the same day, the immediate proximity in time ensures that the patient will provide highly similar responses to both sets of diagnostic questions. In terms of a Venn diagram, the degree of criterion contamination from this methodology could be represented as two circles of roughly similar size that almost fully overlap, though one is offset to the right.

Third, moderately low coefficients would be expected when the criterion diagnoses are derived from an independently conducted structured interview that is completed within the same week (but not the same day) as the CIDI or DIS interview. In terms of a Venn diagram, the circles would still largely overlap but now one would be offset further to the right.

Fourth, low coefficients would be expected when criterion diagnoses systematically draw on a wide range of information, including clinical inferences that emerge from observing the patient over time, from unstructured interviews and interactions, from input by significant others in the patient's life, from input by other mental health professionals who have treated the patient, from historical records, and so on. Following the LEAD model (i.e., Longitudinal, Expert evaluation of All Data; Spitzer, 1983), criterion diagnoses may even use CIDI or DIS information. Although this would produce some degree of criterion contamination or source overlap, it should not be very problematic so long as the CIDI or DIS results were just one of many sources of information that influenced the final diagnostic decisions. In terms of Venn imagery, the CIDI or DIS diagnoses relative to the criterion diagnoses would be like a dime inside or partially overlapping a basketball.

The prior paragraph describes the optimal design for producing independent and unconfounded criterion diagnoses. Unfortunately, few studies meet these conditions. Thus, for the current analysis, the final category contained all remaining studies that did not fit within one of the three categories described earlier. As such, it included criterion diagnoses that were derived from formal team consensus, from traditional clinical interviews, from observations over the course of a hospitalization, from chart reviews, and from independently conducted semistructured clinical interviews. For all of these studies, the criterion diagnoses should have less source contamination with the CIDI or DIS diagnoses than found in the other three designs.

To test these expectations for how methodological overlap influences the apparent accuracy of diagnoses from fully structured interviews, one would need a large number of relevant studies. Fortunately, the literature contains many studies that have examined the correspondence between fully structured interviews like the CIDI or DIS and an external diagnostic criterion. Furthermore, a wide variety of designs have been used in the literature.

After conducting a thorough search of MEDLINE and PsycINFO and following up on all the pertinent citations listed in relevant articles, I found 43 adult samples that examined the validity of fully structured diagnostic interviews. The first column of Table 3.2 reports citation information for these 43 samples. The next seven columns report methodological or design

TABLE 3.2
The Validity of Fully Structured Diagnostic Interviews Relative to Clinical Diagnoses: Studies, Potential Moderators, and Average Agreement

Study	False BR	Lifetime Dx	Crit incl pred	Same int	Same day	Same week	Both struc quest	Expt corr	# Dx	*N*	Mean κ
Anthony et al., 1985	1	0	0	0	0	0	1	1	0	810	.15
Brugha et al., 1999	1	0	0	0	0	0	1	1	15	205	.11
Eaton et al., 2000[a]	1	1	0	0	0	0	1	1	1	349	.26
Erdman et al., 1987	1	0	0	0	0	0	0	1	14	220	.14
Goethe & Ahmadi, 1991[b]	0	0	1	0	0	0	0	1	1	162	.56
Helzer et al., 1985	1	1	0	0	0	0	1	1	11	370	.35
Koenig et al., 1989	1	0	0	0	0	1	0	1	1	69	.38
Mathisen et al., 1987	0	0	0	0	0	1	0	1	5	135	.34
McLeod et al., 1990	0	0	0	0	0	0	1	1	4	347	.21
Murphy et al., 2000	1	1	0	0	0	0	1	1	1	139	.26
North et al., 1997[a]	1	1	0	0	0	0	0	1	5	97	.52
North et al., 1997	1	1	0	0	0	0	0	1	5	33	.37
Peters & Andrews, 1995	0	0	0	0	0	0	0	1	7	98	.36
L. N. Robins et al., 1982[a]	0	0	0	0	0	0	0	1	10	167	.23
Rosenman et al., 1997	0	0	0	0	0	1	0	1	70[c]	126	.10
Thornton, Russell, & Hudson, 1998	1	1	0	0	0	0	1	1	3	44	.60
Anduaga, Forteza, & Lira, 1991	1	1	0	0	0	1	1	2	14	149	.30
Booth, Kirchner, Hamilton, Harrell, & Smith, 1998	1	0	0	0	0	1	1	2	1	54	.53
Compton, Cottler, Dorsey, Spitznagel, & Mager, 1996	1	1	0	0	0	1	1	2	4	123	.57
Cooper, Peters, & Andrews, 1998	1	0	0	0	0	1	1	2	2	73	.13
Cottler et al., 1997; Pull et al., 1997[d]	1	1	0	0	0	1	1	2	12	420	.32
Hesselbrock, Stabenau, Hesselbrock, Mirkin, & Meyer, 1982	1	1	0	0	0	1	1	2	7	42	.80
Hwu, Yeh, & Chang, 1986[a]	0	0	0	0	0	1	1	2	3	100	.69
Ross, Swinson, Larkin, & Doumani, 1994	0	1	0	0	0	1	1	2	12	174	.31
Andrews, Peters, Guzman, & Bird, 1995	0	0	0	0	1	1	1	3	5	101	.40
Burnam, Karno, Hough, Escobar, & Forsythe, 1983	0	1	0	0	1	1	1	3	21	151	.30
Canino et al., 1987[e]	1	1	0	0	1	1	1	3	16	189	.56
Canino et al., 1987[e]	1	1	0	0	1	1	1	3	16	189	.42
Farmer, Katz, McGuffin, & Bobbington, 1987	1	0	0	0	1	1	1	3	5	30	.55
Hwu et al., 1986	1	0	0	0	1	1	1	3	11	187	.42
Peters et al., 1996	1	1	0	0	1	1	1	3	1	91	.46
L. N. Robins et al., 1981	1	1	0	0	1	1	1	3	18	204	.68
Ross et al., 1994[a,f]	0	1	1	0	0	1	1	3	11	80	.51
Wittchen et al., 1985	1	1	0	0	1	1	1	3	12	262	.62
Canino et al., 1987	1	1	1	1	1	1	1	4	16	189	.59
Helzer et al., 1985	1	1	1	1	1	1	1	4	11	370	.79
Janca, Robins, Bucholz, Early, & Shayka, 1992, & Janca, Robins, Cottler, & Early, 1992	0	0	1	1	1	1	1	4	3	20	.79
Kovess, Sylla, Fournier, & Flavigny, 1992	0	1	1	1	1	1	1	4	7	139	.66
Robins, Helzer, Ratcliff, & Seyfried, 1982	1	1	1	1	1	1	1	4	10	204	.65
Wittchen, Kessler, Zhao, & Abelson, 1995	1	1	1	1	1	1	1	4	1	36	.47
Wittchen, Zhao, Abelson, Abelson, & Kessler, 1996	1	1	1	1	1	1	1	4	1	40	.32
Wittchen et al., 1996	1	1	1	1	1	1	1	4	1	37	.54
Wittchen et al., 1996	1	1	1	1	1	1	1	4	1	34	.79
Correlation of moderator with mean κ	.08	.36	.48	.47	.49	.43	.29	.60	−.29	−.31	

Note. BR = base rate; Dx = diagnosis; Crit incl pred = criterion diagnosis included or considered the predictor diagnosis; Int = interview; Both struc quest = both the criterion and predictor diagnoses were derived from a structured set of questions; Expt corr = expected correspondence between predictor and criterion (1 = low, 4 = high); # Dx = number of diagnoses examined. [a]Kappa coefficients were computed or recomputed for this study. [b]Even though Diagnostic Interview Schedule (DIS) diagnoses were available to clinicians when making final diagnoses, this study was classified as having low expected validity because the criterion diagnoses took into account a wide range of additional information. [c]This represents the maximum number of diagnoses that could have been considered. The authors did not indicate how many different diagnoses were assigned by the Composite International Diagnostic Interview (CIDI) or by the clinicians. [d]One of these studies reported correspondence with *Diagnostic and Statistical manual of Mental Disorders* criterion diagnoses, while the other reported correspondence with *International Classification of Diseases* criterion diagnoses. For this table, agreement was averaged across diagnostic systems. [e]These entries use the same sample of patients. The first entry compared lay administered DIS diagnoses to criterion diagnoses made by psychiatrists after they independently completed the DIS. The second entry compared the lay DIS diagnoses to the diagnoses made by the psychiatrists after completing the DIS and any additional follow-up questions they felt were necessary for diagnostic clarification. [f]This study was classified as having moderately high expected validity because DIS diagnoses played a large role in final diagnostic decisions. The latter were made from the results of two Structured Clinical Interview for *DSM-III-R* interviews in conjunction with the DIS findings.

features that were considered potential moderators of validity. These include whether the study used criterion diagnoses that were (a) derived from an artificial base rate (i.e., where diagnostic variance was increased by using diverse targeted patient groups or by mixing nonpatients with targeted patient groups), (b) based on lifetime rather than current symptomatology, (c) made after considering information from the CIDI or DIS interview, (d) made by clinicians who had attended or performed the CIDI or DIS interviews, (e) made on the same day as the CIDI or DIS diagnoses, (f) made during the same week as the CIDI or DIS diagnoses, and (g) derived from a structured set of questions or criteria (like those found on the CIDI and DIS). The ninth column contains the codes for my global expectations regarding source overlap after considering all of the methodological features in each study. In this scheme, 1 indicates the least confounded design, which should generate the lowest validity coefficients, and 4 indicates the most confounded design, which should generate the highest agreement coefficients. The next three columns report the number of diagnoses examined in each sample, the number of participants in the sample, and the average agreement (expressed by the kappa coefficient) between CIDI or DIS diagnoses and criterion diagnoses. Finally, the last row at the bottom of the table reports the correlation between each of the potential moderators and the average agreement rates that are reported in the last column.

Before examining the impact of study design on the magnitude of the final results, it is useful to briefly consider the correlations in the last row of Table 3.2. Contrary to my expectation, an artificial base rate did not have a positive univariate association with the magnitude of agreement. Although I did not have clear expectations for lifetime versus current diagnoses, lifetime diagnoses led to higher rates of diagnostic agreement. The remaining coefficients were statistically significant in the expected direction. Validity coefficients were larger when the criterion judgments took into consideration the predictor results, when criterion diagnoses were made by clinicians who attended or performed the predictor interviews, when both predictor and criterion diagnoses were determined on the same day or in the same week, and when a structured set of items or questions was used to make criterion diagnoses. As expected, the overall ranking of study designs had the largest association with the average study results (r = .60).[2] Finally, both the number of diagnoses under consideration and the number of patients in the study had a negative relationship with the overall magnitude of the findings. Thus, researchers who examined just a few diagnostic conditions in a specialty setting obtained higher estimates of CIDI or DIS validity, whereas those who conducted comprehensive large-scale studies tended to obtain lower estimates of validity.

[2]When considering all potential moderators except for the overall classification scheme, the best predictors of average agreement across studies were the following: (a) criterion diagnoses that were made in the same week as the predictor diagnoses, (b) criterion diagnoses that were made after considering results from the CIDI or DIS predictor, (c) the number of diagnoses that were considered (inversely), and (d) a focus on lifetime rather than current diagnoses. With either a forward stepwise entry model or a backward removal model, the multiple R was .67 for these variables.

Table 3.3 presents summary information concerning the methodological impact of criterion contamination on CIDI or DIS validity. Individual studies were organized according to the four categories of source overlap that were described earlier and listed in column 9 of Table 3.2. Recall that the extent of source overlap should predict the extent of agreement observed in the literature. The table reports the number of studies and participants in each of the four source overlap categories and provides several alternative measures for the average agreement coefficient in the literature. Moving from left to right, the summary coefficients go from the least refined (i.e., the unweighted mean, where every study is given equal weight) to the most refined (i.e., the mean after weighting studies by the number of participants included and the number of diagnoses considered). Considering coefficients in the last column, one can see that methodological artifacts have a very large impact on the apparent extent to which CIDI or DIS diagnoses accurately predict the criterion diagnoses. The studies with the least amount of source overlap and criterion contamination produce the lowest estimates of accuracy (i.e., mean = .19). As predictor diagnoses and criterion diagnoses share more source contamination, the coefficients steadily increase in magnitude (i.e., means = .34 and .51) until they reach their peak, which occurs when clinicians make their criterion diagnoses on the basis of just the CIDI or DIS interview and a limited number of additional, nonscripted follow-up questions (mean = .69). Thus, source overlap plays a mighty role in determining the seeming accuracy of fully structured diagnostic interviews across studies in the research literature.

Next, these data on validity are linked with data on reliability to demonstrate that methodological factors affecting cross-source agreement

TABLE 3.3
Summary of the Accuracy of Fully Structured Diagnostic Interviews as a Function of Predictor–Criterion Overlap

Degree of Source Overlap	K	Total *N*	Mean κ		
			Unweighted	*N* weighted	*N* and # of Dx weighted
Low	16	3,371	.3090	.2510	.1867[a]
Moderately low	8	1,135	.4562	.3898	.3414
Moderately high	10	1,484	.4933	.5067	.5087
High	9	1,069	.6235	.6763	.6903
Across all samples	43	7,059	.4451	.3915	.3689

Note. K = number of studies; Dx = diagnoses. [a]One study considered up to 70 diagnoses, which was an outlier relative to other studies. When the number of diagnoses was limited to a maximum of 30 to make this study more consistent with others, the sample and diagnosis weighted average κ was .2032.

are on a single uniform continuum. The continuum is one that characterizes the independence of information sources. The upper end of this continuum should be anchored by reliability designs that make use of maximally similar information, whereas the lower end should be anchored by validity coefficients that are derived from truly independent sources of information.

More than 40 years ago, Campbell and Fiske (1959) succinctly described this dimension and the terms *reliability* and *validity*. Their classic article contained the section "Convergence of Independent Methods: The Distinction Between Reliability and Validity." In it they said the following:

> Both reliability and validity concepts require that agreement between measures be demonstrated. A common denominator that most validity concepts share in contradistinction to reliability is that this agreement represents the convergence of independent approaches. . . . The importance of independence recurs in most discussions of proof. . . . Independence is, of course, a matter of degree, and in this sense, reliability and validity can be seen as regions on a continuum. . . . Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods. (p. 83)

Campbell and Fisk proceeded to illustrate this continuum by discussing both reliability and validity designs that have greater and lesser degrees of source overlap. Thus, from their perspective, the propensity to think of coefficients in dichotomous terms as indicating either "reliability" or "validity" is quite mistaken. Various types of reliability designs produce a continuum of source independence, as do various types of validity designs. Furthermore, as Campbell and Fiske pointed out, traditional concepts of reliability and validity do not form distinct regions on the source overlap continuum. Some designs that are traditionally considered indicative of validity could be more properly considered indicative of reliability and vice versa.

Previously, I reported CIDI reliability data from Wittchen's (1994) review of the literature. In line with a continuum of source independence, diagnostic decisions derived from a joint interview design should produce the highest reliability coefficients because all information is held constant. Determinations from separate, independently conducted interviews should produce lower reliability coefficients, and coefficients should decline as the delay between interviews becomes longer. Wittchen presented fairly extensive evidence on the reliability of joint interviews. He reported a mean kappa of .92 (N = 575) across 21 diagnostic categories for this type of design. These reliability data were added as the fifth point on the continuum of methodological overlap. Consequently, the continuum had now the following range: 1 = low source overlap and low expected agreement coefficients; 2 = moderately-low source overlap and moderately-low expected agreement coefficients; 3 = moderately-high source overlap and moderately-high expected agreement coefficients; 4 = high source overlap and high expected agreement coefficients; and 5 = total source overlap (i.e., joint interview reliability) and highest expected agreement coefficients.

Figure 3.1 displays the average coefficients relative to each of these five categories. There is a near-perfect linear relationship between these five methodological categories and the average degree of cross-source agreement observed in the research literature. Reliability and validity clearly are on a single continuum. This continuum varies as a function of how much

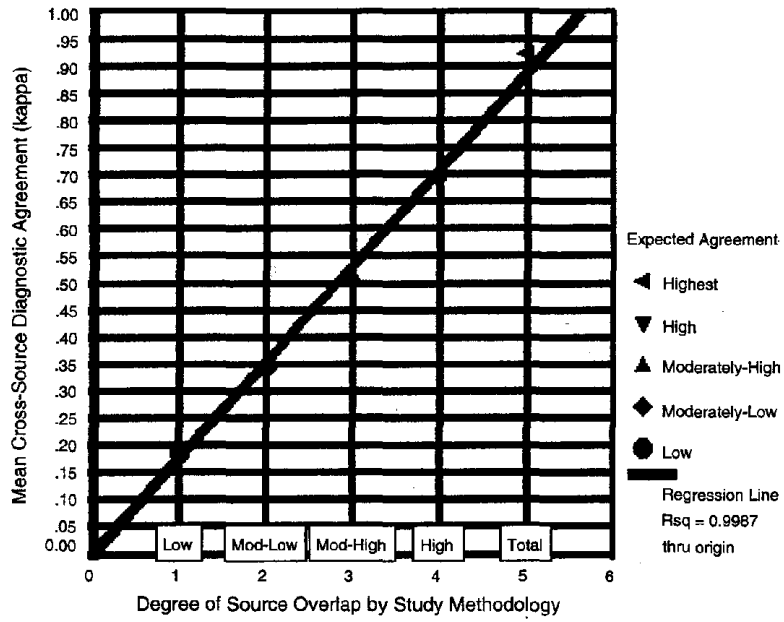

Figure 3.1. Average cross-source agreement for fully structured diagnostic interviews as a function of criterion contamination based on methodologically driven source overlap. The source overlap categories are as follows: Low = predictor diagnoses derived from structured interviews and criterion diagnoses derived from alternative and largely independent sources of information (N = 3,371); Moderately-Low = predictor and criterion diagnoses derived from independently administered structured interviews within the same week (N = 1,135); Moderately-High = predictor and criterion diagnoses derived from independently administered structured interviews on the same day (N = 1,484); High = criterion diagnoses derived from clinicians who considered little information beyond the structured interviews that also determined the predictor diagnoses (N = 1,135); Total = joint interview reliability designs (N = 575). Rsq = R^2.

one source of measurement overlaps or influences the other source of measurement. Thus, the validity or accuracy of a single source of measurement relative to a common criterion is not fixed. Rather, accuracy fluctuates as a function of how much source overlap exists between the predictor and the criterion. In terms of diagnoses, the accuracy of CIDI or DIS diagnoses that are determined by patient self-reports relative to clinician diagnoses is not constant. Indeed, many studies in the research literature produce an illusion of CIDI and DIS accuracy because they use a design that is heavily confounded by source overlap, and this artificially inflates the statistical results.

Although I anticipated a positive relationship as one moved up the validity–reliability continuum, the remarkable linear association in Figure 3.1 was unexpected. After the average coefficients are organized by methodological features in the primary studies, they have virtually no deviation from a perfect linear relationship (a regression line through the origin of Figure 3.1 explains 99.87% of the variance in the aggregated coefficients; i.e., $R^2 = .9987$). However, this result is accidental in the sense that there was no reason to believe that the five gross categories of source overlap on the horizontal axis should have produced essentially equal interval scaling. For instance, if sufficient data from other reliability designs with high source overlap had been available, it is possible that the joint reliability coefficients that are currently in category 5 would have been placed in a category labeled 6 or 7. Doing so would have produced a curvilinear relationship in Figure 3.1.

Similarly, the figure could have plotted results from a reliability design that is relatively free of source overlap problems (e.g., the stability of lifetime disorders over 1 year with baseline and retest interviews conducted by independent clinicians). If one were thinking in terms of traditional reliability versus validity categories rather than in terms of source overlap, the results from these relatively unconfounded reliability studies could have been (erroneously) assigned to a category greater than 4. Doing so would have again produced a different looking graph than that presented in Figure 3.1.

Despite these issues, the linear relationship displayed in Figure 3.1 provides strong support for the overall hypothesis. The figure could also provide a tentative template for examining the results of alternative methodological designs. For instance, if a sample of studies examining the 1-year retest stability of independently conducted CIDI or DIS interviews produced an average agreement coefficient of .40, Figure 3.1 would suggest that these studies have a source overlap "score" of approximately 2.25. In turn, this would indicate that this reliability design has less contamination from source overlap than two of the validity designs commonly reported in the literature.

The data presented in Table 3.3 and Figure 3.1 have been discussed

in terms of the accuracy of CIDI or DIS diagnoses as predictor variables. However, these same data speak to the accuracy of the criterion diagnoses. Every study in Table 3.2 used what the authors believed would be reasonable or optimal clinical diagnoses to validate the CIDI or DIS. If criterion clinical diagnoses were genuine clinical realities, they would remain fixed or constant across studies, and the findings from different samples would converge on the same statistical parameter regarding the accuracy of CIDI or DIS diagnoses. As such, the only factor that would influence the observed estimates of accuracy from one study to the next would be sampling error (e.g., Hunter & Schmidt, 1990). However, because the average agreement coefficients vary widely, the data clearly indicate that the literature is not estimating a single statistical parameter for CIDI and DIS accuracy or for the accuracy of the criterion diagnoses.

Instead, criterion diagnoses are shaped by and conform to the sources of information that are considered in the diagnostic process. When a clinician has only the patient's self-report on a given occasion as a source of information, diagnostic decisions conform to the limitations imposed by that source. As the clinician considers more and more independent sources of information, diagnostic decisions increasingly diverge from the decisions that would be suggested by that self-reported information.

The upshot of this is important. In the absence of conflicting or contradictory data from independent sources, a clinician who consults a single source of information makes decisions that conform to the data provided by that source. The clinician is likely to feel confident about those decisions (after all, there would be no data to suggest otherwise), but as Table 3.3 and Figure 3.1 indicate, these decisions are most likely to be unstable, ephemeral, and ultimately inaccurate.

STRATEGIES FOR SYNTHESIZING INFORMATION AND DATA INDICATING THE VALUE OF DOING SO

The findings presented so far indicate a primary challenge that faces any system developed to replace or supplement the current *DSM*. Regardless of the specific content that is used to define pathology, an empirically guided alternative system should formally require clinicians who use the system to systematically consider multiple, maximally independent sources of information prior to making diagnostic determinations.

If the extensive body of data reviewed in Table 3.1 is taken to heart, it is clear that any single source of information is incomplete and limited. Sophisticated clinicians and researchers should expect associations of about .20–.30 between alternative data sources, and this should fuel the motivation to systematically gather data from multiple independent sources whenever an accurate understanding of the patient is required.

These principles really should not come as a surprise to psychologists. They have been embedded in the classic literature on psychometrics (e.g., Campbell & Fiske, 1959) and experimental design (e.g., Cook & Campbell, 1979). For instance, Cook and Campbell illuminated how the construct validity of nomothetic research can be seriously compromised by mono-method and mono-operation bias. As such, optimal nomothetic research maximizes construct validity by obtaining data from multiple methods of assessment, multiple sources of information, and multiple operational definitions of the target construct. Meyer et al. (1998; Meyer et al., 2001) pointed out how these nomothetic research principles also provide an optimal guide for idiographic clinical practice. Just as optimal research recognizes that any given source of information is incomplete, optimal clinical practice should recognize that identical limitations are in place when conceptualizing the complex life of a single patient. As such, clinicians should seek to minimize mono-method and mono-operation bias in their clinical work by synthesizing information gathered from multiple sources.

Several strategies can be used for synthesizing multisource information. These include simple procedures of mathematical aggregation, the integration of data by a single clinician, and the integration of data by two or more clinicians. In the process of describing the strengths and limitations of these approaches, I also highlight the emerging evidence that indicates these strategies improve the quality of information used in clinical decision making.

Simple Mathematical Aggregation

A long psychometric history speaks to the value of an aggregated mean or sum as a procedure for improving the reliability and validity of information. Indeed, when the principle of aggregation is applied to the number of items in a scale (akin to the number of criteria required for a diagnosis), it forms the basis of the well-known Spearman-Brown Prophecy Formula for estimating the reliability of a composite. (For overviews and recent extensions, see Li, Rosenthal, and Rubin, 1996, and Drewes, 2000.) More important, classics in the research literature have demonstrated how aggregating information over occasions (i.e., longitudinally), over stimuli (e.g., one diagnostic interview format and another), over methods of measurement (e.g., highly structured and unstructured), and over sources of information (e.g., self-report and spouse report) can drastically increase the validity of the aggregated information (see Epstein, 1980, 1983; Rushton, Brainerd, & Pressley, 1983).

Tsujimoto et al. (1990) presented a formula for estimating the validity of a score or judgment that should emerge when aggregating across sources of information. Although Tsujimoto et al.'s formula was devised for use with different kinds of raters, it could also incorporate different occasions,

stimuli, or methods as sources of data. The expected validity of a composite (subject to sampling error) is a function of three things: the degree of validity typically found with a single source of information, the number of sources that contribute data to the aggregated ratings, and the average pairwise agreement between the different sources. Considering this information, Tsujimoto et al. demonstrate how the validity of the aggregated score or judgment hinges on the average agreement between information sources. Although superficially counterintuitive, validity increases when cross-source agreement decreases. Thus, validity is maximized when sources of information provide independent and nonredundant data about the same phenomenon.

How practical is this approach to integrating information? Consider an example related to diagnoses. Consistent with Table 3.1, assume that a diagnostic construct obtained from the patient's self-report correlates with the same construct assessed by an external judge at a magnitude of .30. Further assume that other sources of information (e.g., parent, spouse, friend, coworker, therapist) also agree with each other at a magnitude of .30. Theoretically, according to Tsujimoto et al.'s (1990) formula, if a final diagnostic determination was based on the average rating from three of these sources (e.g., the patient, a peer, and his or her therapist), the expected validity of the diagnostic construct would be .41 rather than .30. This is a salient positive change. Note, however, that this example draws on three independent sources of information with low levels of agreement between them. If the sources were largely redundant, the expected validity would not improve much. For instance, consider a patient's self-reports of diagnostic symptomatology and the diagnostic ratings provided by two clinicians who attended the same highly structured interview with the patient. If we assume that the true validity of the patient's rating remains fixed at .30 and assume that the two clinicians produce ratings that correlate with each other at .80, the validity of the three-rater composite only increases from .30 to .32.

The main advantage of using mathematical aggregation to maximize validity is that it is simple. Sophisticated clinical judgment is not required to synthesize the multisource information. Instead, every data source is treated as equally valid and equally error prone. Thus, all that is required is some mechanism for computing an average or combined score across the different sources. The chief disadvantage of this approach is identical to its main advantage. Because every data source is treated as equally error prone and because clinical judgment is not used to synthesize the information, meaningful and important cross-method discrepancies are likely to be overlooked.

For instance, with this approach, one would not be able to identify the propensity of narcissistic individuals to present themselves in an overly favorable and biased light relative to the view held by external observers

(e.g., Colvin, Block, & Funder, 1995; Gosling, John, Craik, & Robins, 1998; R. W. Robins & John, 1997). Indeed, whereas research has clearly demonstrated that aggregated information provides enhanced validity for assessing the moods and personality traits of nonpatients (e.g., Cheek, 1982; Kolar, Funder, & Colvin, 1996; Moskowitz & Schwarz, 1982; Watson & Clark, 1991), the findings have been mixed in the area of clinical diagnoses (e.g., Bird, Gould, & Staghezza, 1992; Ezpeleta, de la Osa, Doménech, Navarro, & Losilla, 1997; Jensen et al., 1995; Offord et al., 1996; Piacentini, Cohen, & Cohen, 1992; Rubio-Stipec et al., 1994; Schwab-Stone et al., 1996).

For instance, Offord et al. (1996) examined several methods for combining diagnostic information about a large sample of children using data from parents and teachers (N = 1,134). They found that composite information was not decisively more valid than information derived from a single source. In fact, efforts to combine data across sources appeared to mask the distinctive and unique information that could be obtained from each individual source. As a result, Offord et al. concluded that childhood diagnoses should be considered source specific entities. Diagnoses derived from parents are inherently different than diagnoses derived from teachers, and both should be treated as distinct and unique data rather than combined.

Integration of Multisource Information by a Clinician

A clear advantage of having a clinician integrate multisource information is that the clinician can attend to meaningful cross-source discrepancies on the basis of the patient's idiographic characteristics. Thus, unlike the simple mathematical rules described above, a clinician can consider forms of bias that may affect different sources (e.g., the patient's denial mechanisms or excessive criticisms, the observer's inability to access the patient's internal states) or genuine variations in the way problems may be expressed across settings (e.g., that which is seen by a spouse but not a peer or vice versa) in order to identify how the data fit a meaningful pattern or when one source of information should be given more credence than another. When done properly, this would allow the clinician to synthesize all the available multisource information to generate a cohesive and individualized picture of the patient and his or her presenting problems.

The disadvantage of this approach lies in its complexity and the absence of clear models that specify how to weigh opposing information from distinct sources. To meaningfully integrate discrepant sources of data requires sophisticated clinical judgment, a theory about how people are put together, and a capacity to diligently and faithfully think through all of the observed discrepancies. This can be accomplished by clinicians working alone, although it is often a challenge because clinical reasoning is itself subject to various lapses in judgment (see, e.g., Arkes, 1981; Borum, Otto, & Golding, 1993; Garb, 1998).

Integration by Two or More Clinicians

The final option for integrating multisource data is to have two or more clinicians synthesize information from various sources. This approach allows for an idiographic understanding of meaningful cross-method disparities while simultaneously minimizing the errors that can creep into the reasoning of a single clinician working alone. This process can be implemented formally through team discussions or informally through consultation. Much of the relevant research literature has examined the more formal mechanism of team discussion, although this may be too expensive or unrealistic to implement in many settings. As such, informal consultation with another professional can provide a practical alternative. This often can be accomplished most readily by referring patients for a multimethod psychological or neuropsychological evaluation (see Meyer et al., 1998; Meyer et al., 2001).

Probably the most frequently researched procedure for synthesizing multisource information emerges from what Spitzer (1983) briefly described as the longitudinal expert evaluation of all data (LEAD) model for generating gold standard judgments. In this approach, expert clinicians evaluate a patient over time and independently use all available sources of information to formulate a diagnosis or make a description of the patient's personality. Subsequently, the experts jointly review the determinations they had made independently, identify points of disagreement, and resolve any discrepancies to arrive at final decisions. Although time intensive and demanding, the LEAD method has been successfully applied in numerous investigations (e.g., Fennig, Craig, Tanenberg-Karant, & Bromet, 1994; Klein, Ouimette, Kelly, Ferro, & Riso, 1994; Leckman, Sholomskas, Thompson, Belanger, & Weissman, 1982; Mazaide et al., 1992; Pilkonis, Heape, Ruddy, & Serrao, 1991). When properly implemented, the LEAD procedure provides clinical data that come about as close to optimal as possible.

A number of studies have examined the interrater reliability of these best-estimate, multisource diagnoses and have found them to be reproducible despite the absence of explicit guidelines for how clinicians should go about integrating the discrepant information (e.g., Fennig et al., 1994; Leckman et al., 1982; Maziade et al., 1992). However, because the same sources of information are used by all the clinicians making determinations (i.e., interview transcripts, medical records), the design is similar to the joint interview reliability study, in which there is extensive source overlap (e.g., Figure 3.1).

More impressive evidence regarding the value of multisource deter-

minations would emerge from studies that fall further along on the validity end of the continuum. Although I am not aware of studies comparing the relative validity of single-source and multisource diagnostic assessments against a truly external criterion measure (e.g., independently measured etiological factors), studies have examined the stability of diagnoses that should remain constant over time. These include personality disorders and diagnoses that are assigned for the lifetime history of a disorder. If multisource assessments are necessary to accurately capture clinical phenomena, then lifetime disorders and personality disorders should show higher levels of stability when they are diagnosed using multisource procedures rather than single sources of information.

The most compelling evidence on this topic comes from the research of Daniel Klein and his colleagues (1994). These investigators devised a set of seven guidelines to integrate the information provided by patients, relatives, and treatment records in the context of deliberations made by a team of clinicians. They reported retest stability coefficients (kappa) for multisource determinations that were made twice over a period of approximately 2.5 years. Diagnostic determinations were made for 12 personality disorder classifications and for 7 Axis I lifetime diagnoses. Their results are presented in Tables 3.4 and 3.5, along with relevant comparison data.

Table 3.4 deals with the stability of Axis II personality disorders. This table also contains multisource results from a study by Pilkonis et al. (1991). The two multisource assessment studies are presented on the left, and the results from seven single-source assessment studies are on the right. The multisource findings from Klein et al. (1994) in the second column can be compared in a fairly direct manner with the single-source findings that are reported in the fourth column. Both sets of results were generated from the same ongoing project, but the latter coefficients were obtained from the Personality Disorder Examination (PDE), which is a semistructured interview that relies on the patient as the sole source of information (see Ferro, Klein, Schwartz, Kasch, & Leader, 1998). The Pilkonis et al. (1991) study provides a direct comparison for multisource assessments (3rd column) and single-source assessments made with the PDE (8th column). The remaining studies in Table 3.4 provide single-source stability coefficients across time periods that ranged from 2 years to less than 2 weeks.

Overall, Table 3.4 reveals that multisource assessments produce substantially larger stability coefficients than single-source assessments. In fact, all of the 2.5-year stability coefficients from Klein et al.'s study of synthesized judgments exceed the stability coefficients reported in each of the single-source studies, including the very short 1- to 14-day retest coefficients that were obtained by First et al. (1995). Similarly, the multisource determinations from Pilkonis et al. are more stable over 6 months than the relevant comparison data drawn from equivalent or shorter retest in-

TABLE 3.4
The Stability of Personality Disorder Diagnoses Derived From Multiple vs. Single Sources of Information as Measured by the Kappa Coefficient

	Multisource		Single Source						
	Klein[a] (2.5 y, N = 92)	Pilkonis[b] (6 m, N = 31)	Klein[c] (2.5 y, N = 92)	Caccicola[d] (2 y, N = 219)	Mattanah[e] (2 y, N = 65)	Weiss[f] (1 y, N = 31)	Pilkonis[g] (6 m, N = 31)	Zimmerman[h] (5 m, N = 66–457)	First[i] (1–14 d, N = 284)
Any personality disorder	.93	.84	.22	.51	.18		.54	.58	.49
Any Cluster A disorder	.78		.60	.12	.14	.47			.45
Any Cluster C disorder	.91		.34	.47	.20	.27			.56
Any Cluster B disorder	.75		.28	.25	.19	.42			.44
Schizoid	.88		.42[j]	.09					
Antisocial	.94		.16[j]	.45		.41		.48	.71
Borderline	.84		.40	.30	−.01	.02		.61	.48
Histrionic	.86		.48	−.06		.59		.43	.62
Narcissistic	.93		−.02[j]	.35		.59		.25	.42
Avoidant	.82		.33[j]	.23		−.15		.47	.54
Obsessive–compulsive	.73		.10[j]	−.04		.26		.35	.24
Passive–aggressive	.90		.16[j]	.21	.16	.63		.09	.41
Mean	.86	.84	.29	.24	.14	.35	.54	.41	.49

Note. y = years; m = months; d = days. [a]Klein et al., 1994. [b]Pilkonis et al., 1991. [c]Ferro et al., 1998 (this sample was drawn from the same overall project as Klein et al., 1994). [d]Cacciola, Rutherford, Alterman, McKay, & Mulvaney, 1998. [e]Mattanah, Becker, Levy, Edell, & McGlashan, 1995 (this study used adolescents, which may have contributed to lower stability; it also used a quasi-longitudinal expert evaluation of all data procedure at baseline, although the follow-up data were drawn solely from a structured interview with the patient). [f]Weiss, Najavits, Muenz, & Hufford, 1995 (coefficients were computed from raw data). [g]This study uses the same sample as that used in the Pilkonis et al., 1991, multisource diagnoses. [h]Zimmerman, 1994 (standardized interview data were obtained from the review presented in Table 3 and were supplemented with Loranger et al.'s [1994] more detailed results. The specific sample sizes were as follows: any personality disorder, 391; antisocial, 363; borderline, 457; histrionic, 423; narcissistic, 144; avoidant, 421; obsessive–compulsive, 101; and passive–aggressive, 66). [i]First et al., 1995 (when it was not possible to obtain exact coefficients across all participants from their tables, the weighted average coefficient was computed from the patient and nonpatient samples). [j]Coefficients were generated from the reported data, but they were not calculated by the authors because the disorders had a base rate less than .05 at either baseline or retest. Only narcissistic personality disorder had a base rate less than .05 on both occasions.

TABLE 3.5
The Stability of Lifetime Axis I Diagnoses Derived From Multiple vs. Single Sources of Information as Measured by the Kappa Coefficient

Disorder	Klein[a] (2.5 y, *N* = 92)	Rice[b] (6 y, *N* = 1,629)	Prusoff[c] (24.7 y, *N* = 413)	Fendrich[d] (2 y, *N* = 69)	Bromet[e] (1.5 y, *N* = 391)	Keller[f] (6 m, *N* = 84–95)	Andreasen[g] (6 m, *N* = 50)	Helzer[h] (6 w, *N* = 394)	Williams[i] (1–14 days, *N* = 592)
Major depression	.90	.61	.80	.54	.41	.36	.75	.50	.62
Recurrent major depression	.77						.21		
Dysthymia	.75					.53			.44
Alcoholism	.86	.70	.41	.66			.72	.68	.74
Drug abuse	1.00	.56	.56	.66				.52	.85
Phobia	1.00	.34	.32	.33				.36[j]	.50[j]
Any anxiety disorder	.94	.30[k]	.55[k]	.41			.15[k]		.56[k]
Mean	.89	.50	.53	.52	.41	.45	.46	.52	.62

Note. All studies were single source except Klein, which was multisource. [a]Klein et al. (1994). [b]Rice, Rochberg, Endicott, Lavori, and Miller (1992). [c]Prusoff, Merikangas, and Weissman (1988). [d]Fendrich, Weissman, Warner, and Mufson (1990). Coefficients are from participants aged 18 years or older at the baseline evaluation. The coefficient for any substance abuse is reported in both the alcoholism and drug abuse row. [e]Bromet, Dunn, Connell, Dew, and Schulberg (1986). [f]Keller et al. (1995). [g]Andreasen et al. (1981). The follow-up diagnoses in this sample were obtained from the consensus of two independent interviewers. [h]Helzer et al. (1985). [i]Williams et al. (1992). [j]Coefficient reflects the average value for social phobia and simple phobia. [k]Coefficient is for generalized anxiety disorder.

tervals. These findings clearly illustrate the importance of conducting a multisource assessment to maximize the validity of clinical determinations.

Similar findings are evident for the lifetime Axis I disorders presented in Table 3.5. The 2.5-year stability coefficients from Klein et al.'s (1994) multisource, synthesized clinical judgments are presented on the left, and single-source comparison data drawn from various retest intervals are presented on the right. Note that two of the comparison studies used a substantially longer re-evaluation interval than Klein et al. (4.7 and 6 years vs. 2.5 years), so they should be expected to produce lower stability coefficients (e.g., Roberts & DelVecchio, 2000). Moreover, none of the comparison studies draw on the same sample of patients as Klein et al., and the coefficients fluctuate substantially from one sample to another (e.g., the stability of a major depressive disorder is about .60 in the two largest studies, even though the retest interval is just 1–14 days in one study and it is 6 years in the other). Because differences in samples and study designs undoubtedly contribute to the fluctuating stability coefficients, it is more difficult to isolate the impact of a multisource assessment method. Nevertheless, the results appear clear-cut. Determinations about Axis I lifetime disorders that are made on the basis of multiple, independent sources of information appear to be much more valid than the same determinations made from interviews with the patient as the sole source of information.

CONCLUSION

This chapter has a very simple take-home message. If we wish to accurately describe or classify patients, we must synthesize information from multiple independent methods and multiple independent sources. This is not just a good idea. Rather, it is a necessity if we wish to have a refined, empirically grounded taxonomy of psychological health and illness. The data presented here have been collected over many years, across a diverse range of patients, and over a broad domain of psychological attributes. They amply illustrate how any classification system applied to human functioning and pathology remains ephemeral and elusive unless clinical determinations are grounded in carefully synthesized judgments that incorporate data drawn from diverse sources of information, assessment methods, and points in time.

REFERENCES

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213–232.

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

Andreasen, N. C., Grove, W. M., Shapiro, R. W., Keller, M. B., Hirschfeld, R. M. A., & McDonald-Scott, P. (1981). Reliability of lifetime diagnosis: A multicenter collaborative perspective. *Archives of General Psychiatry, 38*, 400–405

Andrews, G., Peters, L., Guzman, A. M., & Bird, K. (1995). A comparison of two structured diagnostic interviews: CIDI and SCAN. *Australian and New Zealand Journal of Psychiatry, 29*, 124–132.

Anduaga, J. C., Forteza, C. G., & Lira, L. R. (1991). Concurrent validity of the DIS: Experience with psychiatric patients in Mexico City. *Hispanic Journal of Behavioral Sciences, 13*, 63–77.

Anthony, J. C., Folstein, M., Romanoski, A. J., Von Korff, M. R., Nestadt, G. R., Chahal, R., Merchant, A., Brown, C. H., Shapiro, S., Kramer, M., & Gruenberg, E. M. (1985). Comparison of the lay Diagnostic Interview Schedule and a standardized psychiatric diagnosis: Experience in Eastern Baltimore. *Archives of General Psychiatry, 42*, 667–675.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49*, 323–330.

Bernstein, D. P., Kasapis, C., Bergman, A., Weld, E., Mitropoulou, V., Horvath, T., Klar, H., Silverman, J., & Siever, L. J. (1997). Assessing Axis II disorders by informant interview. *Journal of Personality Disorders, 11*, 158–167.

Bird, H. R., Gould, M. S., & Staghezza, B. (1992). Aggregating data from multiple informants in child psychiatry epidemiological research. *Journal of the American Academy of Child and Adolescent Psychiatry, 31*, 78–85.

Booth, B. M., Kirchner, J. E., Hamilton, G., Harrell, R., & Smith, G. R. (1998). Diagnosing depression in the medically ill: Validity of a lay-administered structured diagnostic interview. *Journal of Psychiatric Research, 32*, 353–360.

Borum, R., Otto, R., & Golding, S. (1993). Improving clinical judgment and decision making in forensic evaluation. *Journal of Psychiatry & Law, 21*, 35–76.

Bromet, E. J., Dunn, L. O., Connell, M. M., Dew, M. A., & Schulberg, H. C. (1986). Long-term reliability of diagnosing lifetime major depression in a community sample. *Archives of General Psychiatry, 43*, 435–440.

Brugha, T. S., Bebbington, P. E., Jenkins, R., Meltzer, H., Taub, N. A., Jana, M., & Vernon, J. (1999). Cross validation of a general population survey diagnostic interview: A comparison of CIS–R with SCAN *ICD–10* diagnostic categories. *Psychological Medicine, 29*, 1029–1042.

Burnam, M. A., Karno, M., Hough, R. L., Escobar, J. I., & Forsythe, A. B. (1983). The Spanish Diagnostic Interview Schedule: Reliability and comparison with clinical diagnoses. *Archives of General Psychiatry, 40*, 1189–1196.

Cacciola, J. S., Rutherford, M. J., Alterman, A. I., McKay, J. R., & Mulvaney, F. D. (1998). Long-term test–retest reliability of personality disorder diagnoses in opiate dependent patients. *Journal of Personality Disorders, 12*, 332–337.

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

Canino, G. J., Bird, H. R., Shrout, P. E., Rubio-Stipec, M., Bravo, M., Martinez, R., Sesman, M., Guzman, A., Guevara, L. M., & Costas, H. (1987). The Spanish Diagnostic Interview Schedule: Reliability and concordance with clinical diagnoses in Puerto Rico. *Archives of General Psychiatry, 44*, 720–726.

Cheek, J. M. (1982). Aggregation, moderator variables, and the validity of personality tests: A peer-rating study. *Journal of Personality and Social Psychology, 43*, 1254–1269.

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

Compton, W. M., Cottler, L. B., Dorsey, K. B., Spitznagel, E. L., & Mager, D. E. (1996). Comparing assessments of *DSM–IV* substance dependence disorders using CIDI–SAM and SCAN. *Drug and Alcohol Dependence, 41*, 179–187.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Cooper, L., Peters, L., & Andrews, G. (1998). Validity of the Composite International Diagnostic Interview (CIDI) psychosis module in a psychiatric setting. *Journal of Psychiatric Research, 32*, 361–368.

Cottler, L. B., Grant, B. F., Blaine, J., Mavreas, V., Pull, C., Hasin, D., Compton, W. M., Rubio-Stipec, M., & Mager, D. (1997). Concordance of *DSM–IV* alcohol and drug use disorder criteria and diagnoses as measured by AUDADIS–ADR, CIDI, and SCAN. *Drug and Alcohol Dependence, 47*, 195–205.

Drewes, D. W. (2000). Beyond Spearman-Brown: A structural approach to maximal reliability. *Psychological Methods, 5*, 214–227.

Eaton, W. W., Neufeld, K., Chen, L.-S., & Cai, G. (2000). A comparison of self-report and clinical diagnostic interviews for depression: Diagnostic Interview Schedule and Schedules for Clinical Assessment in Neuropsychiatry in the Baltimore Epidemiologic Catchment Area follow-up. *Archives of General Psychiatry, 57*, 217–222.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist, 35*, 790–806.

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

Erdman, H. P., Klein, M. H., Greist, J. H., Bass, S. M., Bires, J. K., & Machtinger, P. E. (1987). A comparison of the Diagnostic Interview Schedule and clinical diagnosis. *American Journal of Psychiatry, 144*, 1477–1480.

Ezpeleta, L., de la Osa, N., Doménech, J. M., Navarro, J. B., & Losilla, J. M. (1997). Diagnostic agreement between clinicians and the Diagnostic Interview for Children and Adolescents (DICA–R) in an outpatient sample. *Journal of Child Psychology and Psychiatry, 38*, 431–440.

Farmer, A. E., Katz, R., McGuffin, P., & Bobbington, P. (1987). A comparison

between the Present State Examination and the Composite International Diagnostic Interview. *Archives of General Psychiatry, 44*, 1064–1068.

Fendrich, M., Weissman, M. M., Warner, V., & Mufson, L. (1990). Two-year recall of lifetime diagnoses in offspring at high and low risk for major depression: The stability of offspring reports. *Archives of General Psychiatry, 47*, 1121–1127.

Fennig, S., Craig, T. J., Tanenberg-Karant, M., & Bromet, E. J. (1994). Comparison of facility and research diagnoses in first-admission psychotic patients. *American Journal of Psychiatry, 151*, 1423–1429.

Ferro, T., Klein, D. N., Schwartz, J. E., Kasch, K. L., & Leader, J. B. (1998). 30-month stability of personality disorder diagnoses in depressed outpatients. *American Journal of Psychiatry, 155*, 653–659.

First, M. B., Spitzer, R. L., Gibbon, M., Williams, J. B. W., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope, H. G., Jr., & Rounsaville, B. (1995). The Structured Clinical Interview for *DSM–III–R* Personality Disorders (SCID–II). Part II: Multi-site test-retest reliability study. *Journal of Personality Disorders, 9*, 92–104.

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

Goethe, J. W., & Ahmadi, K. S. (1991). Comparison of Diagnostic Interview Schedule to psychiatrist diagnoses of alcohol use disorder in psychiatric inpatients. *American Journal of Drug and Alcohol Abuse, 17*, 61–69.

Gómez-Beneyto, M., Villar, M., Renovell, M., Pérez, F., Hernandez, M., Leal, C., Cuquerella, M., Slok, C., & Asencio, A. (1994). The diagnosis of personality disorder with a modified version of the SCID–II in a Spanish clinical sample. *Journal of Personality Disorders, 8*, 104–110.

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line coding of observers. *Journal of Personality and Social Psychology, 74*, 1337–1349.

Helzer, J. E., Robins, L. N., McEvoy, L. T., Spitznagel, E. L., Stoltzman, R. K., Farmer, A., & Brockington, I. F. (1985). A comparison of clinical and Diagnostic Interview Schedule diagnoses: Physician reexamination of lay-interviewed cases in the general population. *Archives of General Psychiatry, 42*, 657–666.

Hesselbrock, V., Stabenau, J., Hesselbrock, M., Mirkin, P., & Meyer, R. (1982). A comparison of two interview schedules: The Schedule for Affective Disorders and Schizophrenia-Lifetime and the National Institute for Mental Health Diagnostic Interview Schedule. *Archives of General Psychiatry, 39*, 674–677.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hwu, H. G., Yeh, E. K., & Chang, L. Y. (1986). Chinese Diagnostic Interview Schedule: I. Agreement with psychiatrist's diagnosis. *Acta Psychiatrica Scandinavica, 73*, 225–233.

Janca, A., Robins, L. N., Bucholz, K. K., Early, T. S., & Shayka, J. J. (1992). Comparison of Composite International Diagnostic Interview and clinical *DSM–III–R* criteria checklist diagnoses. *Acta Psychiatrica Scandinavica, 85*, 440–443.

Janca, A., Robins, L. N., Cottler, L. B., & Early, T. S. (1992). Clinical observation of assessment using the Composite International Diagnostic Interview (CIDI): An analysis of the CIDI Field Trials–Wave II at the St Louis site. *British Journal of Psychiatry, 160*, 815–818.

Jensen, P., Roper, M., Fisher, P., Piacentini, J., Canino, G., Richters, J., Rubio-Stipec, M., Dulcan, M., Goodman, S., Davies, M., Rae, D., Shaffer, D., Bird, H., Lahey, B., & Schwab-Stone, M. (1995). Test–retest reliability of the Diagnostic Interview Schedule for Children (DISC 2.1). *Archives of General Psychiatry, 52*, 61–71.

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.

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

Keller, M. B., Klein, D. N., Hirschfeld, R. M. A., Kocsis, J. H., McCullough, J. P., Miller, I., First, M. B., Holzer, C. P., III, Keitner, G. I., Marin, D. B., & Shea, T. (1995). Results of the *DSM–IV* mood disorders field trial. *American Journal of Psychiatry, 152*, 843–849.

Klein, D. N., Ouimette, P. C., Kelly, H. S., Ferro, T., & Riso, L. P. (1994). Test–retest reliability of team consensus best-estimate diagnoses of Axis I and II disorders in a family study. *American Journal of Psychiatry, 151*, 1043–1047.

Koenig, H. G., Goli, V., Shelp, F., Cohen, H. J., Meador, K. G., & Blazer, D. G. (1989). Major depression and the NIMH Diagnostic Interview Schedule: Validation in medically ill hospitalized patients. *International Journal of Psychiatry and Medicine, 19*, 123–132.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64*, 311–337.

Kovess, V., Sylla, O., Fournier, L., & Flavigny, V. (1992). Why discrepancies exist between structured diagnostic interviews and clinicians' diagnoses. *Social Psychiatry and Psychiatric Epidemiology, 27*, 185–191.

Leckman, J. F., Sholomskas, D., Thompson, W. D., Belanger, A., & Weissman, M. M. (1982). Best estimate of lifetime psychiatric diagnosis. *Archives of General Psychiatry, 39*, 879–883.

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods, 1*, 98–107.

Loranger, A. W., Sartorius, N., Andreoli, A., Berger, P., Buchheim, P., Channabasavanna, S. M., Coid, B., Dahl, A., Diekstra, R. F. W., Ferguson, B., Jacobsberg, L. B., Mombour, W., Pull, C., Ono, Y., & Regier, D. A. (1994).

The International Personality Disorder Examination: The World Health Organization/Alcohol, Drug Abuse, and Mental Health Administration international pilot study of personality disorders. *Archives of General Psychiatry, 51*, 215–224.

Maffei, C., Fossati, A., Agostoni, I., Barraco, A., Bagnato, M., Deborah, D., Namia, C., Novella, L., & Petrachi, M. (1997). Interrater reliability and internal consistency of the Structured Clinical Interview for *DSM–IV* Axis II Personality Disorders (SCID–II), Version 2.0. *Journal of Personality Disorders, 11*, 279–284.

Malgady, R. G., Rogler, L. H., & Tryon, W. W. (1992). Issues of validity in the Diagnostic Interview Schedule. *Journal of Psychiatric Research, 26*, 59–67.

Mathisen, K. S., Evans, F. J., & Meyers, K. (1987). Evaluation of a computerized version of the Diagnostic Interview Schedule. *Hospital and Community Psychiatry, 38*, 1311–1315.

Mattanah, J. J. F., Becker, D. F., Levy, K. N., Edel, W. S., & McGlashan, T. H. (1995). Diagnostic stability in adolescents followed up 2 years after hospitalization. *American Journal of Psychiatry, 152*, 889–894.

Maziade, M., Roy, M.-A., Fournier, J.-P., Cliche, D., Mérette, C., Caron, C., Garneau, Y., Montgrain, N., Shriqui, C., Dion, C., Nicole, L., Potvin, A., Lavallée, J.-C., Pirès, A., & Raymond, V. (1992). Reliability of best-estimate diagnosis in genetic linkage studies of major psychoses: Results from the Quebec Pedigree Studies. *American Journal of Psychiatry, 149*, 1674–1686.

McLeod, J. D., Turnbull, J. E., Kessler, R. C., & Abelson, J. M. (1990). Sources of discrepancy in the comparison of a lay-administered diagnostic instrument with clinical diagnoses. *Psychiatric Research, 31*, 145–159.

Meyer, G. J. (1997). On the integration of personality assessment methods: The Rorschach and MMPI-2. *Journal of Personality Assessment, 68*, 297–330.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., Eisman, E. J., & Dies, R. R. (1998). *Benefits and costs of psychological assessment in healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group, Part I.* Washington, DC: American Psychological Association.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128–165.

Michell, J. (1997). Quantitative science and the definition of *measurement* in psychology. *British Journal of Psychology, 88*, 355–383.

Mischel, W. (1996). *Personality and assessment.* Mahwah, NJ: Erlbaum. (Original work published 1968)

Moskowitz, D. S., & Schwarz, J. C. (1982). Validity comparison of behavior counts and ratings by knowledgeable informants. *Journal of Personality and Social Psychology, 42*, 518–528.

Murphy, J. M., Monson, R. R., Laird, N. M., Sobol, A. M., & Leighton, A. H. (2000). A comparison of diagnostic interviews for depression in the Stirling County study: Challenges for psychiatric epidemiology. *Archives of General Psychiatry, 57*, 230–236.

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

North, C. S., Pollio, D. E., Thompson, S. J., Ricci, D. A., Smith, E. M., & Spitznagel, E. L. (1997). A comparison of clinical and structured interview diagnoses in a homeless mental health clinic. *Community Mental Health Journal, 33*, 531–543.

Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., & Lipman, E. L. (1996). Integrating data from multiple informants. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*, 1078–1085.

Peters, L., & Andrews, G. (1995). Procedural validity of the computerized version of the Composite International Diagnostic Interview (CIDI–Auto) in the anxiety disorders. *Psychological Medicine, 25*, 1269–1280.

Peters, L., Andrews, G., Cottler, L. B., Chatterji, S., Janca, A., & Smeets, R. M. W. (1996). The Composite International Diagnostic Interview posttraumatic stress disorder module: Preliminary data. *International Journal of Methods in Psychiatric Research, 6*, 167–174.

Piacentini, J. C., Cohen, P., & Cohen, J. (1992). Combining discrepant diagnostic information from multiple sources: Are complex algorithms better than simple ones? *Journal of Abnormal Child Psychology, 20*, 51–63.

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, 3*, 46–54.

Prusoff, B. A., Merikangas, K. R., & Weissman, M. M. (1988). Lifetime prevalence and age of onset of psychiatric disorders: Recall 4 years later. *Journal of Psychiatric Research, 22*, 107–117.

Pull, C. B., Saunders, J. B., Mavreas, V., Cottler, L. B., Grant, B. F., Hasin, D. S., Blaine, J., Mager, D., & Üstün, B. T. (1997). Concordance between *ICD–10* alcohol and drug use disorder criteria and diagnoses as measured by the AUDADIS–ADR, CIDI and SCAN: Results of a cross-national study. *Drug and Alcohol Dependence, 47*, 207–216.

Rice, J. P., Rochberg, N., Endicott, J., Lavori, P. W., & Miller, C. (1992). Stability of psychiatric diagnoses: An application to the affective disorders. *Archives of General Psychiatry, 49*, 824–830.

Richter, P., Werner, J., Heerlein, A., Kraus, A., & Sauer, H. (1998). On the validity of the Beck Depression Inventory. A review. *Psychopathology, 31*, 160–168.

Riskind, J. H., Beck, A. T., Berchick, R. J., Brown, G., & Steer, R. A. (1987). Reliability of *DSM–III* diagnoses for major depression and generalized anxiety disorder using the Structured Clinical Interview for *DSM–III*. *Archives of General Psychiatry, 44*, 817–820.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of per-

sonality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126,* 3–25.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: Its history, characteristics, and validity. *Archives of General Psychiatry, 38,* 381–389.

Robins, L. N., Helzer, J. E., Ratcliff, K. S., & Seyfried, W. (1982). Validity of the Diagnostic Interview Schedule, Version II: *DSM–III* diagnoses. *Psychological Medicine, 12,* 855–870.

Robins, R. W., & John, O. P. (1997). Effects of visual perspective and narcissism on self-perception: Is seeing believing? *Psychological Science, 8,* 37–42.

Rogler, L. H., Malgady, R. G., & Tryon, W. W. (1992). Evaluation of mental health: Issues of memory in the Diagnostic Interview Schedule. *Journal of Nervous and Mental Disease, 180,* 215–222.

Rosenman, S. J., Korten, A. E., & Levings, C. T. (1997). Computerized diagnosis in acute psychiatry: Validity of CIDI–Auto against routine clinical diagnosis. *Journal of Psychiatric Research, 31,* 581–592.

Ross, H. E., Swinson, R., Larkin, E. J., & Doumani, S. (1994). Diagnosing comorbidity in substance abusers: Computer assessment and clinical validation. *Journal of Nervous and Mental Disease, 182,* 556–563.

Rubio-Stipec, M., Canino, G. J., Shrout, P., Dulcan, M., Freeman, D., & Bravo, M. (1994). Psychometric properties of parents and children as informants in child psychiatry epidemiology with the Spanish Diagnostic Interview Schedule for Children (DISC.2). *Journal of Abnormal Child Psychology, 22,* 703–720.

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

Schwab-Stone, M. E., Shaffer, D., Dulcan, M. K., Jensen, P. S., Fisher, P., Bird, H. R., Goodman, S. H., Lahey, B. B., Lichtman, J. H., Canino, G., Rubio-Stipec, M., & Rae, D. S. (1996). Criterion validity of the NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC–2.3). *Journal of the American Academy of Child and Adolescent Psychiatry, 35,* 878–888.

Segal, D. L., Hersen, M., & Van Hasselt, V. B. (1994). Reliability of the Structured Clinical Interview for *DSM–III–R*: An evaluative review. *Comprehensive Psychiatry, 35,* 316–327.

Skre, I., Onstad, S., Torgersen, S., & Kringlen, E. (1991). High interrater reliability for the Structured Clinical Interview for *DSM–III–R* Axis I (SCID–I). *Acta Psychiatrica Scandinavica, 84,* 167–173.

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

Spitzer, R. L., & Fleiss, J. L. (1974). A re-analysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry, 125,* 341–347.

Spitzer, R. L., Forman, J. B. W., & Nee, J. (1979). *DSM–III* field trials: I. Initial interrater diagnostic reliability. *American Journal of Psychiatry, 136,* 815–817.

Steinberg, M., Rounsaville, B., & Cicchetti, D. V. (1990). The Structured Clinical Interview for *DSM–III–R* Dissociative Disorders: Preliminary report on a new diagnostic instrument. *American Journal of Psychiatry, 147,* 76–82.

Thornton, C., Russell, J., & Hudson, J. (1998). Does the Composite International Diagnostic Interview underdiagnose the eating disorders? *International Journal of Eating Disorders, 23,* 341–345.

Tsujimoto, R. N., Hamilton, M., & Berger, D. E. (1990). Averaging multiple judges to improve validity: Aid to planning cost-effective clinical research. *Psychological Assessment, 2,* 432–437.

Watson, D., & Clark, L. A. (1991). Self- versus peer ratings of specific emotional traits: Evidence of convergent and discriminant validity. *Journal of Personality and Social Psychology, 60,* 927–940.

Weiss, R. D., Najavits, L. M., Muenz, L. R., & Hufford, C. (1995). Twelve-month test–retest reliability of the structured clinical interview for *DSM–III–R* personality disorders in cocaine dependent patients. *Comprehensive Psychiatry, 36,* 384–389.

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.

Williams, J. B. W., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope, H. G., Jr., Rounsaville, B., & Wittchen, H.-U. (1992). The Structured Clinical Interview for *DSM–III–R* (SCID): II. Multisite test–retest reliability. *Archives of General Psychiatry, 49,* 630–636.

Wittchen, H.-U. (1994). Reliability and validity studies of the WHO–Composite International Diagnostic Interview (CIDI): A critical review. *Journal of Psychiatric Research, 28,* 57–84.

Wittchen, H.-U., Kessler, R. C., Zhao, S., & Abelson, J. (1995). Reliability and clinical validity of UM–CIDI *DSM–III–R* generalized anxiety disorder. *Journal of Psychiatric Research, 29,* 95–110.

Wittchen, H.-U., Semler, G., & von Zerssen, D. (1985). A comparison of two diagnostic methods: Clinical *ICD* diagnoses vs *DSM–III* and Research Diagnostic Criteria using the Diagnostic Interview Schedule (version 2). *Archives of General Psychiatry, 42,* 677–684.

Wittchen, H.-U., Zhao, S., Abelson, J. M., Abelson, J. L., & Kessler, R. C. (1996). Reliability and procedural validity of UM–CIDI *DSM–III–R* phobic disorders. *Psychological Medicine, 26,* 1169–1177.

Zimmerman, M. (1994). Diagnosing personality disorders: A review of issues and research methods. *Archives of General Psychiatry, 51,* 225–245.