Journal of Consulting and Clinical Psychology
2002, Vol. 70, No. 1, 158–168

0022-006X/02/$5.00 DOI: 10.1037//0022-006X.70.1.158

Assessing Match and Mismatch Between Practitioner-Generated and Standardized Interview-Generated Diagnoses for Clinic-Referred Children and Adolescents

Amanda L. Jensen and John R. Weisz
University of California, Los Angeles

Although the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) is widely used in both clinical and research settings, little is known about agreement between clinician and standardized research diagnoses. Diagnoses generated by the Diagnostic Interview Schedule for Children (DISC-P–2.3) were compared with clinician-generated diagnoses for 245 referred youths. Agreement was poor for all individual disorders and broader diagnostic clusters. Agreement was higher for externalizing categories than for internalizing, but no association was found between agreement and child, family, or clinician characteristics. Clinicians were more likely than the DISC to assign 1 diagnosis and less likely to assign 0 diagnoses, suggesting that clinic policies may play a role. Implications for the use of the *DSM* across different settings are discussed.

Over the past 30 years, categorical diagnostic classification has become standard practice in mental health care. In the United States, this trend is reflected in the widespread use of the *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed. [*DSM–III*], 3rd ed., rev. [*DSM–III–R*], and 4th ed. [*DSM–IV*]; American Psychiatric Association, 1980, 1987, 1994) in virtually all clinical service settings. A formal *DSM* diagnosis is now required for admission and treatment in most mental health facilities and programs, for most third-party reimbursement of diagnostic and treatment services, and for a variety of other kinds of authorization, intervention, and formal record keeping. In addition, the diagnoses generated for this purpose also figure prominently in treatment planning, as clinicians shape interventions to address the diagnoses assigned.

The framers of the *DSM* have long recognized that the system will be used in both clinical practice and research. The introduction to *DSM–III–R*, for example, states that a purpose of the manual is "to provide clear descriptions of diagnostic categories in order to enable clinicians and investigators to diagnose, communicate about, study and treat the various mental disorders" (American Psychiatric Association, 1987, p. xxix). Although the present study focuses on *DSM–III–R*, similar general points were made in the introductions to the *DSM–III* (American Psychiatric Association, 1980, p. 12) and *DSM–IV* (American Psychiatric Association, 1994, p. xv). Although research and clinical applications of the *DSM* system are compatible in many respects, there are some notable differences between research and clinical settings in the conditions under which *DSM* assessment is carried out and *DSM* diagnoses are determined and assigned. For example, in diagnostic assessment done for research purposes (e.g., epidemiological studies and clinical trials), research interviewers trained to use standardized procedures apply those procedures to study volunteers who have agreed (and are often paid) to participate in interviews as long as it takes to permit coverage of the full diagnostic system (often as long as 2–3 hr) or the portions relevant to the study. Often, the purpose of these research interviews is to generate a relatively complete diagnostic profile for the participants, so diagnostic comprehensiveness is often a goal. In clinical diagnostic assessment, by contrast, a major purpose of interviewing is often to identify issues that need to be addressed in treatment, so clinicians may focus on primary problems rather than inquiring about secondary problems that may not be part of the treatment agenda. Moreover, the interviewing is typically done by heavily scheduled clinical staff who have not been given standardized training and do not use standardized interview procedures. The interviewees are clinical clients who are paying (either personally or through third-party payers) for services. Also, the time pressures brought on by cost- and productivity-conscious clinic policies typically limit the time that can be devoted to diagnostic interviewing, so that only some portions of the *DSM*'s child-relevant taxonomy can be covered. In these settings, the fact that a diagnosis is required for authorization of services and/or reimbursement may create an incentive to assign at least one diagnosis; but workload and time constraints may create a disincentive to do comprehensive assessment and assign multiple diagnoses.

For children,[1] clinic versus research differences in length and comprehensiveness may be substantial, given the size and com-

Amanda L. Jensen, Department of Psychology, University of California, Los Angeles (UCLA); John R. Weisz, Department of Psychology and Department of Psychiatry and Biobehavioral Sciences, UCLA.

The study was supported by National Institute of Mental Health (NIMH) Grant R01 MH 49522 and by NIMH Senior Research Scientist Award K05 MH01161, which we gratefully acknowledge. We thank Howard Adelman, Kristin Hawley, John Piacentini, and Tom Wickens for their assistance in various aspects of the study, as well as the participating children, parents, and clinic staff members for their investment of time and thought.

Correspondence concerning this article should be addressed to Amanda L. Jensen or John R. Weisz, Department of Psychology, Franz Hall, University of California, 405 Hilgard Avenue, Los Angeles, California 90095-1563.

[1] We use the term *children* to refer to both children and adolescents, except when we need to distinguish between the two age groups.

plexity of the *DSM* system that is relevant to this age group. The large number of diagnostic categories, with multiple symptoms per diagnosis, means that standardization of interviewer questions is almost certainly needed to ensure comprehensive coverage and that a standardized and comprehensive diagnostic assessment is almost certain to be quite lengthy. This is true of all of the standardized approaches developed thus far (e.g., Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders—V; [SCID; Spitzer, Williams, Gibbon & First, 1992], Diagnostic Interview Schedule [DIS; Robins, Helzer, Croughan & Ratcliff, 1981], Schedule for Affective Disorders and Schizophrenia for School Aged Children [K-SADS; Puig-Antich & Chambers, 1978], Diagnostic Interview for Children and Adolescents [DICA; Herjanic, Herjanic, Brown, & Wheatt, 1975], and Diagnostic Interview Schedule for Children [DISC; Shaffer et al., 1993]). If lengthy standardized interviews are required to fully assess for the *DSM* diagnoses, and if such interviews are characteristic of research diagnosis but not clinical diagnosis, then there is a chance that the end-products of diagnostic assessment (i.e., the diagnoses obtained) in these two kinds of contexts may be rather different from one another even for the same children interviewed at similar points in time. If standardized interview diagnoses do in fact differ substantially from diagnoses generated for the same individuals through typical clinical procedures, such differences would have several implications, to which we now turn.

First, such differences might raise questions about the clinical utility of the diagnostic system, under current conditions of clinical practice. The first goal listed by the *DSM–III–R* framers was "clinical usefulness for making treatment and management decisions in varied clinical settings" (American Psychiatric Association, 1987, p. xix.; see p. xv of *DSM–IV* for statement of a similar goal for that system). Clearly, this is an important goal. A failure of practice-generated diagnoses to match those yielded by more standardized and comprehensive procedures would raise questions about whether the goal has been achieved, that is, whether the diagnostic system is in fact producing accurate treatment and management decisions in clinical settings.

Second, a mismatch between clinician-generated diagnoses and those derived from standardized procedures could raise a question as to whether another of the *DSM–III–R* framers' goals was fully attained, that is, to "enable clinicians and investigators to diagnose, communicate about, study, and treat the various mental disorders" (American Psychiatric Association, 1987 p. xxix). The use of the same diagnostic descriptions and criteria almost certainly seems to have improved communication between these two groups; a common vocabulary is a valuable step toward good communication. However, if the two groups' use of the system leads to divergent judgments as to which diagnoses should be assigned to which individuals, it could be argued that at least some significant communication problems persist.

Third, a mismatch between standardized and clinical diagnoses would have implications for the dissemination of treatment research to practice settings. Over the past decade, task forces within the American Psychological Association (Chambless et al., 1998; Lonigan, Elbert, & Johnson, 1998) have compiled lists of empirically supported treatments and have encouraged their use in clinical practice. Many of these treatments have, of course, been developed and tested with participants identified through standardized interview procedures. If these standardized procedures generate substantially different diagnoses than the participating children would have received through usual clinic procedures, then the children studied in the research may not be very well matched to children actually treated in the clinic. For example, if an empirically tested treatment program for children with depressive disorders were adopted by a mental health clinic for use with its referred youths, the fit of the treatment to the children might not be ideal if youngsters diagnosed with depressive disorders by clinic staff would not have been given those diagnoses by the research team who designed and tested the program.

Is there a mismatch between standardized and clinic diagnoses? Definitive evidence is lacking thus far, but some initial studies offer suggestive findings. In these studies, agreement between clinician diagnoses and those generated with standardized research instruments[2] has generally been assessed by using J. Cohen's (1960) kappa. Following Fleiss (1981), kappas below .40 reflect "poor" agreement, kappas between .40 and .74 reflect "fair to good" agreement, and kappas .75 and higher reflect "excellent" agreement. In our review of the studies, we note kappas where available, and we also note the degree to which clinician diagnoses obtained appear to have represented usual clinical practice. We stress that we are not criticizing the studies as their methods often fit their goals; our intent is simply to note the extent to which their procedures reflected usual clinical practice, an issue of special interest in the present study.

Ezpeleta et al. (1997) compared the Spanish version of the revised version of the Diagnostic Interview for Children and Adolescents (DICA–R; Reich, Shayka, & Taibleson, 1991) with clinician diagnoses from three outpatient clinics; clinicians' diagnoses were produced by means of a checklist provided by the researchers. Kappas ranged from –.04 to 1.00 for the child version and from .07 to .55 for the adolescent version, with most kappas falling in the poor range ($M = .31$). Kappas obtained for the parent versions were higher, ranging from –.02 to 1.00, with most falling in the fair-to-good range ($M = .41$). However, because clinicians generated their diagnoses using a researcher-devised checklist, the clinician diagnoses may not have been very representative of usual clinical practice.

Pellegrino, Singh, and Carmanico (1999) compared child DISC (Shaffer, Fisher, Piacentini, Schwab-Stone, & Wicks, 1989) diagnoses with inpatient and partial hospitalization clinician diagnoses from a single clinic; clinicians had been instructed to use their usual diagnostic procedures but were then provided with a list of the DISC-generated diagnoses to rate whether the children met criteria for any of those disorders. Even guided by this list, which directed attention to the DISC diagnoses, clinicians showed generally poor agreement with DISC diagnoses; kappas ranged from .03 to .61 ($M = .22$). Prompting clinicians with a list of DISC-

[2] In Schwab-Stone et al. (1996), clinician diagnoses were based on an abbreviated version of the DISC, administered by clinicians, and any further inquiries the clinicians felt were necessary at the end of the structured interview. Piacentini et al. (1993) used a semistructured instrument to generate the clinician diagnoses. And Fisher et al. (1993) compared DISC diagnoses with diagnoses assigned in clinics selected for their expertise in certain diagnoses (e.g., Tic Disorders Clinic at the Yale Child Study Center). All three studies used procedures quite appropriate to their goals, but none used procedures representative of usual practice in most outpatient clinics, which is the focus of the present study.

generated diagnoses does not appear to represent usual clinical practice.

Vitiello, Malone, Buschle, Delaney, and Behar (1990) compared diagnoses generated by the DICA (Herjanic & Reich, 1982) with diagnoses assigned by supervised psychiatric fellows in an inpatient unit in a single hospital. They obtained kappas ranging from –.03 to .62 for the child report version of the DICA ($M = .28$) and from .10 to .48 for the parent report version ($M = .28$). Welner, Reich, Herjanic, Jung, and Amado (1987) compared the child-report version of the DICA (DICA–C; Herjanic et al., 1975) with clinicians' inpatient discharge diagnoses from a single clinic and obtained kappas ranging from –.18 to .52 ($M = .26$). This comparison was also made using the child version of the DISC (DISC–C; Costello, Edelbrock, & Costello, 1985) in two studies based on the same inpatient sample from a single teaching hospital. Weinstein, Stone, Noam, Grimes, and Schwab-Stone (1989) compared child DISC diagnoses with clinicians' admission diagnoses and found kappas in the poor category, ranging from .03 to .17 ($M = .10$). Aronen, Noam, and Weinstein (1993) later compared the same DISC diagnoses with the clinicians' discharge diagnoses and found similarly poor kappas, ranging from –.07 to .22 ($M = .11$). Procedures in these four studies appear to have represented usual clinic diagnostic practice for inpatient settings; not represented in these studies is outpatient practice, which accounts for the great majority of real-world clinical diagnosis, and in which diagnostic agreement may be more difficult to obtain, given the relative brevity of contact between staff and children.

In the present study, we sought to provide a particularly representative assessment of diagnostic agreement, focusing on usual clinical practice in outpatient child clinics as compared with the most widely used standardized diagnostic interview for children. Procedures included a number of strengths manifest separately but not combined in previous studies (reviewed earlier) plus additional features designed to enrich the research. We (a) used an outpatient sample to reflect the majority of child mental health service users; (b) included multiple clinics to increase generalizability and reduce the risk of findings that might reflect idiosyncratic characteristics or practices of a single setting; (c) used clinician diagnoses generated routinely by practitioners who were not given any instructions, information, diagnosis checklists, or special diagnostic training by the research project, so as to represent usual practice and not to artificially inflate agreement; (d) used research diagnoses generated by trained interviewers who had no access to clinician diagnoses and no information beyond the research interview, also to avoid artificially inflating agreement; and (e) used the DISC, the most widely used and thoroughly researched of all the standardized child diagnostic interviews. As best we can determine, the present study is the first to study clinician diagnoses obtained under fully representative, usual child outpatient practice conditions and to cleanly separate clinician diagnosis from standardized assessment. Thus, it is the first study to directly address the specific issues of interest in this particular investigation.

Another design feature differed importantly from previous work. Previous studies of agreement using the DISC have all used the child-report version of the instrument. This may not be ideal, because DISC child-report diagnoses have been shown to be considerably less reliable than parent-report DISC diagnoses (Schwab-Stone et al., 1993), and low reliability places limits on the level of agreement that can be obtained between research and clinician diagnoses. Three previous studies have, in fact, found better agreement between researcher and clinician diagnoses for the parent-report versions than the child-report versions of the DICA (Ezpeleta et al., 1997) and the DISC (Fisher et al., 1993; Piacentini et al., 1993; Schwab-Stone et al., 1996). Such findings suggest that parent reports may be needed for the fairest test of agreement. We followed this procedure in the present study, but we included some child-report components to assess whether agreement changed when child informants were included.

In addition to assessing overall agreement, we explored factors associated with agreement. We tested whether disruptive and externalizing diagnoses would be more likely to show significant clinician-research agreement than the less-outwardly observable internalizing diagnoses. We also assessed the role of child and family characteristics found to be associated with at least some kinds of agreement (e.g., parent–child and interrater) in past research: child age (e.g., Edelbrock, Costello, Dulcan, Conover, & Kala, 1986; Rapee, Barrett, Dadds, & Evans, 1994; Stavrakaki, Vargo, Roberts, & Boodoosingh, 1987), child gender (e.g., Rapee et al., 1994; Verhulst & Van der Ende, 1991), ethnicity (Strakowski et al, 1997), parental depression (e.g., Boyle & Pickles, 1997; Conrad & Hammen, 1989; Fergusson, Lynskey, & Horwood, 1993), and overall parental psychopathology (Chilcoat & Breslau, 1997). We also explored clinician factors (e.g., experience and academic degree) that could potentially be related to agreement.

Finally, our data set allowed us to test possible explanations for any differences found between clinician- and DISC-generated diagnoses, tests generally not found in previous studies on this issue. We investigated, for example, whether agreement was affected by combining clinicians' *rule-out* diagnoses (i.e., diagnoses that they did not feel they could definitively assign at intake) with their assigned diagnoses, to see if the clinicians' ability to defer diagnosis affected agreement. And to investigate the possible impact of clinic time pressure plus requiring at least one diagnosis for authorization of services, we assessed whether the clinicians were more likely than the DISC to assign a single diagnosis and less likely to assign zero diagnoses and multiple diagnoses.

Method

Participants

Participants were families of children ages 7–17, seeking treatment in five outpatient community mental health clinics in southern California. The research assessment took place as soon as it could be scheduled after the clinic intake appointment ($M = 23.4$ days after intake) and occurred before the participants attended any therapy sessions. Participants' clinic records were later accessed to obtain the formal intake diagnoses assigned by the clinical staff. Data collection spanned 1991–1996. Children whose intakes occurred after *DSM–IV* was published (in 1994) were only included in the study if their clinicians specifically indicated in the charts that they used *DSM–III–R* to generate their diagnoses.

The sample of 245 included 163 boys and 82 girls age 7–17 years ($M = 11.13$, $SD = 2.57$); 51% were Caucasian, 15% African American, 15% Latino, 2% Asian/Pacific Islander, and the remainder (17%) mixed or other ethnicity. Fifty-four percent of the parents had completed at least 1 year of college, and mean annual income was $20,410 ($Mdn = \$15,324$, $SD = \$15,689$).

Measures

Diagnostic Interview Schedule for Children. In the research assessment, child diagnoses were obtained using the DISC–2.3 (Shaffer et al., 1993), administered by clinical psychology graduate students who had all received the standard DISC interviewer training program designed by the DISC development team at Columbia University. The DISC is a highly structured interview consisting of a series of yes/no questions. The interviewer's task was to ask the questions verbatim and record the yes/no answers. The answers were then entered into a computer algorithm to generate diagnoses based on *DSM–III–R* criteria (American Psychiatric Association, 1987) for the primary diagnostic categories relevant to children (i.e., anxiety disorders, affective disorders, disruptive behavior disorders, eating disorders, elimination disorders, and tic disorders).

Test–retest reliability for the parent version of the DISC has been shown to be good to excellent, with retest correlations ranging from .77 to .87 (Schwab-Stone et al., 1993). Interrater reliability has also been shown to be high, with intraclass correlations (ICC) of 1.00 for all of the major diagnoses except major depressive episode, which showed ICC of .66 (Shaffer et al., 1993). Investigations have also established criterion validity of the DISC when compared with other structured or semistructured interviews (e.g., Schwab-Stone et al., 1996) and "expert clinician" consensus diagnoses (Fisher et al, 1993).

Of the adults interviewed, 90% were the child's mother figure, 7% the child's father figure, 2% another female relative, and 1% a female nonrelative. Because child-report DISC diagnoses, unlike parent-report diagnoses (see below), have shown only fair to poor test–retest reliability for several diagnostic categories (Schwab-Stone et al., 1993), and poor reliability limits the level of researcher–clinician agreement that can be obtained (see introduction), we relied mainly on the parent-report DISC. However, we also used the child-report DISC for diagnostic categories for which children might be particularly appropriate informants (i.e., depressive disorders and conduct disorder).

Clinician-generated diagnoses and clinician characteristics. Chart diagnoses were the *DSM–III–R* diagnoses generated for each child by the intake clinician assigned to that child, as a part of normal clinic intake, assessment, and treatment planning procedures. As some clinicians diagnosed more than one child, there were a total of 65 therapists in the sample. The clinicians were 68% female; 51% were psychologists, 32% social workers, and 17% other and unspecified (on a professional information form). Some 25% had doctorates, 54% master's degrees, 19% other (e.g., supervised graduate trainees), and 2% unspecified. Information regarding years of experience was available for 26 therapists, who averaged 7.42 years of experience ($SD = 10.63$).

Brief Symptom Inventory (BSI). Information on parental psychopathology was obtained with the BSI (Derogatis, 1992), a 53-item self-report measure, scored and profiled on nine primary symptom dimensions and three global indices of distress. Here we used the total symptom count as an index of overall parental psychopathology and the depression dimension score as the index of parental depression. The BSI has good internal consistency, good test–retest reliability, with an overall stability coefficient of .90, and good discriminant and convergent validity (Derogatis, 1992). In the present sample, alphas were .88 for the depression dimension and .97 for the total scale, and retest correlations were .63 for the depression dimension and .63 for the total scale.

Results

We used Cohen's kappa (J. Cohen, 1960) to assess concordance between DISC and chart diagnoses. Given the number of kappas calculated, alpha for significance tests was set at .01 to reduce risk of chance findings. Kappas were interpreted using Fleiss's (1981) conventions (see introduction).

Specific Diagnoses Generated by Clinicians and DISC: Number and Agreement

The number of positive diagnoses for each of the sample disorders from the DISC and from the charts is shown in Table 1. The DISC generated more diagnoses per child than did clinicians ($M = 2.44$ vs. 1.22, $p < .01$), with only diagnoses of dysthymia and oppositional defiant disorder assigned at a higher rate by clinicians. McNemar's (1947) test showed significant differences ($p < .01$) on 12 of the diagnostic categories (marked by asterisks in Table 1), all showing DISC rates higher than clinician rates.

Because kappas are sensitive to low base rate (Spitznagel & Helzer, 1985), only diagnoses with more than 20 positive cases identified by either source were included in our kappa calculations assessing clinician–DISC agreement. Those diagnoses are shown in Table 2. Kappa values were in the poor range for all these disorders (see Table 2) and were significantly different from zero only for attention-deficit/hyperactivity disorder (ADHD; $p < .01$) and conduct disorder ($p < .001$). For all the remaining disorders, the nonsignificant kappas indicate that the match between DISC and clinicians was no better than chance expectancy.

There were no positive clinician diagnoses for generalized anxiety disorder (GAD) and social phobia, although according to the DISC, 10% of the sample had a diagnosis of GAD and 20% had social phobia. Although the *DSM–III–R* section on anxiety disorders does allow these diagnoses to be assigned to children, the section on anxiety disorders of childhood or adolescence indicates that GAD is the adult version of overanxious disorder (OAD) and

Table 1
No. of Participants and Percentage of Sample With Positive DISC and Chart Diagnoses

Diagnosis	Positive DISC diagnoses		Positive chart diagnoses	
	n	%	*n*	%
Attention-deficit/hyperactivity disorder	87[a]	34.9	23	10.0
Agoraphobia	10[a]	4.0	0	0.0
Avoidant disorder of childhood or adolescence	11[a]	4.4	1	0.4
Conduct disorder	40[a]	16.5	23	9.2
Dysthymia	43	17.2	57	22.9
Elective mutism	1	0.4	1	0.4
Encopresis	9	3.6	5	2.0
Enuresis	18[a]	7.3	5	2.0
Generalized anxiety disorder	23[a]	9.2	0	0.0
Major depressive disorder	40[a]	16.1	3	1.2
Obsessive–compulsive disorder	7	2.8	2	0.8
Oppositional defiant disorder	105	42.6	121	48.6
Overanxious disorder	37[a]	14.9	11	4.4
Panic	0	0.0	1	0.4
Separation anxiety disorder	36[a]	14.5	6	2.4
Social phobia	47[a]	19.3	0	0.0
Specific phobia	41[a]	16.9	1	0.4
Tics	24[a]	10.0	0	0.0
Total	579		260	

Note. DISC = Diagnostic Interview Schedule for Children–2.3, parent version (Shaffer et al., 1993).
[a] Diagnostic frequencies differed significantly, $p < .01$.

Table 2
Kappa Coefficients for Clinician and DISC Diagnoses

Diagnosis	Chart+/DISC+	Chart−/DISC−	Chart+/DISC−	Chart−/DISC+	κ
ADHD	15	144	8	72	.142*
Dysthymia	9	153	48	34	−.026
Conduct disorder	11	189	11	29	.268*
Oppositional defiant disorder	60	76	59	45	.132
Generalized anxiety disorder	0	221	0	23	.000
Overanxious disorder	3	199	8	34	.060
Major depressive disorder	1	202	2	39	.024
Separation anxiety disorder	1	203	5	35	.006
Simple phobia	1	202	0	40	.040
Social phobia	0	198	0	47	.000

Note. DISC = Diagnostic Interview Schedule for Children–2.3, parent version (Shaffer et al., 1993); Chart = chart diagnosis; + = positive diagnosis; − = negative diagnosis; ADHD = attention-deficit/hyperactivity disorder.
* $p = .01$.

social phobia is the adult version of avoidant disorder of childhood or adolescence. So, given the possibility that the clinicians were conceptualizing the disorders in this manner, we combined the DISC diagnoses of GAD and OAD, and the DISC diagnoses of social phobia and avoidant disorder, and the combined diagnoses were compared with the chart diagnoses of OAD and avoidant disorder, respectively. These analyses, too, yielded nonsignificant kappas in the poor range for both OAD/GAD (.042) and social phobia/avoidant disorder (.033).

Analyses of Broad Diagnostic Clusters

To generate a more inclusive analysis, we also assessed concordance at the level of broad clusters of related diagnoses. We sought to learn the extent to which, even if they disagreed on specific diagnoses, the clinicians and DISC interviews generated a similar picture of the "general idea" or broad domain of the children's problems. Diagnoses were grouped into the following six clusters: Depression, Anxiety, Conduct, Attention/Hyperactivity, Tics, and Elimination. Table 3 shows the specific diagnoses included in each cluster and the number of positive DISC and chart diagnoses for these clusters. McNemar's (1947) tests indicated that the DISC generated significantly more diagnoses for the Anxiety, Attention/Hyperactivity, Tics, and Elimination clusters, whereas clinicians assigned significantly more diagnoses in the Conduct cluster (all ps $< .01$). Kappas were significantly different from zero for the Attention/Hyperactivity, Conduct, and Elimination clusters only ($p < .01$) and fell in the poor range for all clusters except Elimination, which fell in the fair-to-good range (see Table 4).

Factors Associated With Agreement

As outlined in the introduction, we explored three types of variables that might be associated with agreement: problem type,

Table 3
Specific Diagnoses Included in the Five Diagnostic Clusters

Depression (DISC $n = 57$, Chart $n = 60$)	Anxiety (DISC $n = 108$,[a] Chart $n = 22$)	Conduct (DISC $n = 113$,[a] Chart $n = 141$)	Attention/hyperactivity (DISC $n = 87$,[a] Chart $n = 23$)	Tics (DISC $n = 24$,[a] Chart $n = 0$)	Elimination (DISC $n = 24$,[a] Chart $n = 10$)
1. Dysthymia	1. Overanxious disorder of childhood	1. Oppositional defiant disorder	1. Attention-deficit/ hyperactivity disorder	1. Chronic motor tics	1. Encopresis
2. Major depressive disorder	2. Obsessive–compulsive disorder	2. Conduct disorder		2. Chronic vocal tics	2. Enuresis
	3. Elective mutism			3. Tourette's disorder	
	4. Generalized anxiety disorder			4. Transient tic disorder	
	5. Panic disorder				
	6. Social phobia				
	7. Simple phobia				
	8. Separation anxiety				
	9. Agoraphobia				
	10. Avoidant disorder of childhood				

Note. DISC n = number of positive diagnoses generated by the Diagnostic Interview Schedule for Children–2.3, parent version (Shaffer et al., 1993); Chart n = number of positive chart diagnoses.
[a] Diagnostic frequencies differed significantly, $p < .01$.

Table 4
Kappa Coefficients for Clinician and DISC Diagnostic Clusters

Diagnosis	Chart+/ DISC+	Chart−/ DISC−	Chart+/ DISC−	Chart−/ DISC+	κ
Attention/hyperactivity	15	144	8	72	.142*
Depression	14	141	46	43	.000
Anxiety	15	120	7	93	.089
Elimination	8	213	2	16	.437*
Conduct	79	64	62	34	.204*
Tics	0	217	0	24	.000

Note. Chart = chart diagnosis; DISC = Diagnostic Interview Schedule for Children–2.3, parent version (Shaffer et al., 1993); + = positive diagnosis; − = negative diagnosis.
* $p < .01$.

child and family factors, and clinician factors. Each type required a somewhat different kind of test.

Problem type: Externalizing and disruptive disorders versus internalizing disorders. Because there is no accepted method for testing the statistical significance of differences between kappas within the same sample, our test of the effects of problem type on agreement took the form of examining whether, as anticipated, the externalizing, disruptive disorders were more likely than the internalizing disorders to show clinician–DISC agreement significantly higher than chance levels. In analyses at the level of specific diagnoses (see Table 2), kappas had significantly exceeded chance only for conduct disorder ($\kappa = .268$, $p < .01$) and ADHD ($\kappa = .142$, $p < .01$). In the analyses of diagnostic clusters (see Table 4), kappas exceeded chance only for the conduct cluster ($\kappa = .204$, $p < .01$; i.e., conduct disorder and oppositional defiant disorder), the attention/hyperactivity cluster ($\kappa = .142$, $p < .01$), and the elimination disorders cluster ($\kappa = .437$, $p < .01$). Agreement failed to exceed chance for any internalizing diagnosis or cluster.

Child and family factors. To assess for an association between agreement and child or family factors, we computed the proportion of diagnoses agreed upon by both sources for each child by dividing the number of positive diagnoses that the two sources agreed on for that child by the total number of positive diagnoses assigned to the child by either source, creating a variable ranging from 0 (*no diagnostic agreement*) to 1 (*complete diagnostic agreement*). Relationships between agreement and the continuous variables (i.e., child age, family income, parental overall psychopathology, and parental depression) were assessed by calculating the correlation between each of these variables, on the one hand, and the diagnostic agreement index, on the other; this yielded nonsignificant results for all the variables. To determine whether agreement was associated with the categorical variables (i.e., child gender or ethnicity: Caucasian, African American, or Latino), we performed one-way analyses of variance (ANOVAs), with the agreement index as the dependent variable and the child factors as the independent variables. These ANOVAs also yielded nonsignificant results.

Clinician factors. For each clinician, we calculated the proportion of assigned diagnoses agreeing with the DISC by dividing the number of positive diagnoses that the two sources agreed on by the total number of positive diagnoses assigned by either source to the children each clinician worked with, creating a variable that ranged from 0 (*no diagnostic agreement*) to 1 (*complete diagnostic agreement*). To assess the effects of clinician experience on diagnostic agreement, we then correlated this agreement index with clinician years of experience; this yielded nonsignificant results. To determine if there was a relationship between clinician gender, degree (doctoral, master's, other), or professional discipline (social work, psychology, other) and the diagnostic agreement index, we performed one-way ANOVAs, with the agreement index as the dependent variable and the clinician factors as the independent variables. None of these ANOVAs was significant, indicating that clinicians differing on the above variables did not differ in the degree to which their diagnoses were concordant with the DISC diagnoses.

Secondary Analyses to Examine Possible Alternative Explanations for Low Levels of Agreement

One potential explanation for the low levels of agreement we found is that clinicians have the option of identifying rule-out diagnoses for disorders that they might not be comfortable assigning at intake. It may be that, for some disorders, clinicians feel that more extensive assessment is necessary and thus defer assigning the diagnosis until a later date. Although any adjustments the clinicians may have made to the diagnoses at a later date were not available, the rule-out diagnoses were. To assess the validity of this alternative explanation of our findings, we reanalyzed the data with clinician rule-out diagnoses combined with assigned diagnoses, and these combined diagnoses were then compared with the DISC diagnoses. Rule-out diagnoses had been assigned for conduct disorder, ADHD, dysthymia, and major depressive disorder. Inclusion of these rule-out diagnoses had negligible effects on agreement in both the specific disorder analyses and the analyses of diagnostic clusters.

Another potential explanation for the low levels of agreement is that the research diagnoses were based on parent reports alone, whereas clinician diagnoses were based on interviews that sometimes included parents and children. To explore this possible explanation of the findings, we used the child-report DISC diagnoses we had obtained for major depression, dysthymia, and conduct disorder. To test the hypothesis that clinician diagnoses differed from our DISC diagnoses because clinicians had incorporated child reports into their diagnoses, we combined parent and child DISC diagnoses; each child was assigned a positive diagnosis if the parent, the child, or both endorsed the disorder and was

assigned a negative diagnosis only if both failed to endorse the disorder. This method of combining reports has been found to be as effective as more complex statistical approaches (Bird, Gould, & Staghezza, 1992). These new combined diagnoses were then compared with the clinician diagnoses using kappa. Because combined diagnoses were available for both major depression and dysthymia, it was possible to also compare combined diagnoses for the depression cluster. Combining parent and child reports resulted in slightly increased kappas for conduct disorder (from .268 to .378) and for the depression cluster (from .007 to .073) but not for major depression or dysthymia. However, even with the slight increases for conduct disorder and the depression cluster, kappas remained in the poor category, indicating that incorporating child reports did not appreciably increase concordance between the DISC and clinicians.

A third potential explanation for the low agreement was that there was a time lag between the clinical intake and the DISC assessment ($M = 23.44$ days). To test this possibility, we created two groups of children: a *short-lag* group (those whose lag fell in the bottom third of days between the two assessments; $M = 7.12$ days, $n = 79$) and a *long-lag* group (top third; $M = 47.27$ days, $n = 80$). Mean levels of diagnostic agreement for the groups were computed by averaging the diagnostic agreement indices for the children in each group. A *t* test comparing diagnostic agreement for the short-lag and long-lag groups (*M*s = .18 and .22) was not significant; overall agreement was not affected as time between the two assessments increased. We then tested the possibility that the time lag may have affected agreement for some diagnostic categories but not others. For each diagnostic category, a chi-square test was used to compare the short-lag and long-lag groups on the number of children for whom the two sources agreed. None of these tests were significant, indicating that agreement on specific diagnostic categories was also not affected as time between the two assessments increased.

A fourth factor that may have contributed to the low levels of agreement is that in most clinics, including our study clinics, a diagnosis was required for authorization of services; this may have created an incentive for the clinicians to assign at least one diagnosis, an incentive absent from the DISC assessment. To assess whether incentive to assign at least one diagnosis might have affected agreement, we used McNemar's (1947) test to test the hypothesis that the DISC would identified children as having zero diagnoses more often than would clinicians. The hypothesis was supported: The DISC identified 50 children as having no diagnosis, but clinicians only identified 1 child ($p < .001$). Clinicians' workload and time pressures may also have created a disincentive to do comprehensive assessment once the one-diagnosis requirement for service authorization had been met, whereas the DISC interview was designed for comprehensive diagnostic assessment. We used McNemar's test to test the hypothesis that clinicians were more likely than the DISC to assign only one diagnosis per child. The hypothesis was supported: Clinicians assigned a single diagnosis to 149 children and the DISC assigned a single diagnosis to only 60 children ($p < .001$).

Another explanation for the low levels of agreement and the discrepancy in number of diagnoses assigned by the two sources is that clinicians were less likely to detect comorbid conditions (perhaps because of lack of time or a tendency of parents to not provide information about comorbid conditions unless specifically asked). The fact that clinicians were more likely than the DISC to assign only one diagnosis is certainly consistent with this hypothesis. However, it is also possible that different types of comorbidity had differential impact. For example, clinicians may be less likely to note an internalizing disorder once they have identified an externalizing disorder, and vice versa, whereas the DISC may not show such a pattern because it systematically assesses for all diagnoses. To examine patterns of comorbidity within DISC and clinician diagnoses, we computed odds ratios between the different categories of diagnoses for the two sources separately. For the DISC diagnoses, most odds ratios for pairings of diagnoses were significantly larger than 1 ($p < .05$), indicating a positive relationship between the two categories. (The only exceptions to this pattern: Elimination disorders were not significantly associated with any other categories, and anxiety disorders were not significantly associated with tic or conduct disorders.) For the clinician diagnoses, though, an interesting pattern emerged. For all pairings of internalizing disorders with externalizing disorders (e.g., anxiety and ADHD; conduct disorder and depression), the odds ratios were significantly smaller than 1 ($p < .05$), indicating that having a diagnosis of one type meant a reduced likelihood that there would be a diagnosis of the other type. However, for pairings involving *like* disorders (i.e.., ADHD and conduct disorder; anxiety and depression), the odds ratios were not significant, indicating that the presence of an internalizing or externalizing diagnosis did not change the likelihood of a diagnosis of the same type. These findings indicate that clinicians were particularly unlikely to assign an externalizing diagnosis and an internalizing diagnosis to the same child.

Discussion

The findings reveal very low levels of agreement between the diagnoses generated by clinicians in usual clinical practice and the diagnoses generated by standardized interview procedures. Comparisons of clinician-generated and DISC-generated diagnoses using kappa showed consistently poor agreement (Fleiss, 1981). Indeed, for all but two of the diagnoses (conduct disorder and ADHD), agreement levels failed to even exceed chance expectancy. Agreement was not appreciably affected by inclusion of clinicians' rule-out diagnoses or by combining parent and child DISC reports for the depressive disorders and conduct disorder. When we loosened the criteria overall by assessing agreement at the level of broader diagnostic clusters, kappa levels were still poor for all clusters except elimination disorders. Thus, the findings showing poor agreement were robust across multiple methodological variations.

Because there is no gold standard for valid diagnosis, it is impossible to determine whether the DISC- or clinician-based diagnoses are more accurate reflections of true psychopathology in this sample. However, it is useful to consider several possible reasons for the low level of agreement obtained here. First, agreement was likely to have been affected by the fact that the DISC generated more diagnoses than did clinicians for most diagnoses and diagnostic clusters. This difference in diagnostic frequency, in turn, may have multiple explanations. One possibility is that the conditions under which real-world clinicians work do not support the kind of systematic survey of all the main child-relevant *DSM* disorders that is a part of such standardized interviews as the

DISC. Clinic pressures and productivity requirements may limit the time clinicians can devote to diagnostic interviewing, and there may not be opportunities for training in standardized procedures, in any event. Also, as we noted earlier, the purposes of research and clinical diagnostic interviews may differ substantially, with the former aimed at comprehensive diagnostic picture and the latter aimed at identifying issues for treatment; one consequence may be that more diagnoses are generated through research diagnostic interviews than through clinical diagnostic interviews. Finally, although the assignment of at least one diagnosis is often required for approval and reimbursement of services, there may be little incentive to assign additional diagnoses once this requirement has been satisfied. This hypothesis was supported by our finding that the DISC was significantly more likely than clinicians to assign zero diagnoses ($N = 50$ vs. 1) and that clinicians were significantly more likely than the DISC to assign a single diagnosis ($N = 149$ vs. 60).

While the fact that the DISC assigned more diagnoses than clinicians certainly affected agreement between the two, this does not adequately explain all of the disagreement. For example, if different diagnostic frequency fully accounted for the low levels of agreement, disagreement would be highest for those disorders for which the DISC assigned more diagnoses than the clinicians. However, given that agreement in the present study is as low or lower for disorders and clusters that were diagnosed at the same or higher rates in the charts compared with the DISC (e.g., dysthymia, oppositional defiant disorder, and depression cluster) as those diagnosed at a higher rate on the DISC (e.g., ADHD and conduct disorder), overendorsement on the DISC or underendorsement by clinicians cannot adequately explain all of the disagreement. Thus, we turn our attention to other possible explanations for the low agreement.

One possibility is that DISC procedures may be less clinically sensitive than interviews in clinics. First, because administration of the DISC does not allow for clarification of the questions, parents may have incorrectly responded to questions on the DISC because they did not understand them and thus may have overendorsed some symptoms (see Breslau, 1987). Another contributing factor may be that the *DSM–III–R*, and hence the corresponding DISC, used impairment criteria for some but not all symptoms. That is, for only some of the symptoms of the child-relevant diagnoses does *DSM–III–R* require that a symptom be counted only if it is impairing to the child. This may lead to increased numbers of positive diagnoses with standardized procedures that follow strict *DSM* criteria, relative to clinician diagnoses, if clinicians weight impairment more heavily in their interviews. Schwab-Stone et al. (1996) did find the number of DISC diagnoses to be lower, and concordance between the DISC and clinicians employed by the research project and deemed to have good knowledge of *DSM–III–R* to be slightly higher, when the DISC was scored incorporating additional questions concerning child impairment. It is possible that, as the *DSM* system evolves and makes increasing use of impairment criteria, the total number of DISC diagnoses may drop, and agreement between clinicians and standardized interviews may increase somewhat.

Another possibility is that clinicians may have assigned the chart diagnoses used in this study on the basis of their intake interviews but later updated these diagnoses on the basis of additional information gathered as treatment proceeded. Information on subsequent diagnoses that may have been assigned was not available for the present sample, so we were not able to test the hypothesis that, given more exposure to children through therapy, clinicians might assign diagnoses similar to those assigned by the DISC. This hypothesis warrants attention in the future; however, there are reasons to suspect that analyses may not support it. First, note that in our procedures, DISC diagnoses, like clinician diagnoses, were based on interviews carried out prior to therapy sessions; changing the procedure to target clinician diagnoses made *after* therapy session would reduce the parallelism, and thus might actually *reduce* agreement with DISC diagnoses. Second, we were able to test whether incorporation of clinicians' rule-out diagnoses (i.e., the diagnoses they suspected at intake but deferred a decision on, pending further information) increased agreement with the DISC; it did not. And finally, when Aronen et al. (1993) tested whether clinicians' final diagnostic impressions would show better agreement with the DISC than the intake impressions had, they found no increase in agreement. Thus, although we could not directly test whether agreement would increase if clinicians had increased exposure to children through therapy, there are some reasons to doubt that this would be the case.

We examined other potential explanations for low clinician–DISC agreement related to our study procedures. One such procedural factor was the time lag between clinic intake and the DISC assessment ($M = 23.44$ days). Although no intervention occurred during this time lapse, it is possible that during this time, the children's symptoms changed. However, this is not likely to have been a major source of low agreement, given that agreement was poor for such chronic conditions as ADHD and tic disorders. Also, our analyses indicated that neither overall agreement nor agreement on specific diagnostic categories was significantly better for children for whom very little time had passed between the two assessments than for the children for whom much more time had passed, indicating that the time lapse did not affect diagnostic agreement.

An additional potential source of low concordance was that the clinicians had access to both parents and children when making their diagnoses, whereas the DISC diagnoses were based on parent report only. Although parents and children typically show very low levels of agreement with one another (Achenbach, McConaughy, & Howell, 1987), measure validation studies have differed as to whether combining parent and child reports leads to increased concordance between "expert" or research clinicians and standardized interviews (e.g., Fisher et al., 1993; Piacentini et al, 1993) or no increase in concordance (Ezpeleta et al., 1997). Our analyses indicated that combining child and parent reports did lead to somewhat increased agreement for conduct disorder and for the depression cluster. However, even with the inclusion of child reports, agreement for these disorders did not exceed the poor level, indicating that, for these disorders, sole reliance on parent reports did not sufficiently explain the low levels of clinician–DISC agreement. Future research could further assess whether including both parent and child reports improves agreement for other diagnoses.

Another approach to understanding our findings on agreement is to identify factors associated with level of agreement. No relationships were found for child or family factors, including child age, gender or ethnicity, family income, parental overall psychopathology, or parental depression, or for therapist factors, including

gender, years of experience, degree, or professional discipline. Information on one therapist factor that may have made a difference, amount of training and experience with the *DSM*, was not available; however, whatever the average level of such training and experience among our clinicians, it was probably reasonably representative of the level of training and experience found among many clinicians working in community settings and was clearly not sufficient to produce good agreement with standardized diagnostic procedures.

Our clearest finding regarding factors associated with agreement concerned problem type. A few of the most disruptive, externalizing disorders and clusters generated above-chance agreement, whereas no internalizing disorders or clusters did so. This suggests that diagnostic agreement between clinicians and research procedures may be enhanced by salience and observability of child problems, and conversely that agreement may be more difficult to achieve on symptoms that are subtle and partially hidden from view (as in depressive and anxiety disorders).

The identification of problem type as being related to agreement is reminiscent of previous findings suggesting that different childhood problems evoke varying amounts of concern in parents. When Weisz and Weiss (1991) studied the "referability" of child problem behaviors, they found that among American families, undercontrolled problems result in clinic referrals more often than would be expected given their prevalence rates in the population and that overcontrolled problems are referred less often than would be expected from their prevalence rates. Those findings suggest that parents are more likely to seek treatment for problems such as aggression than for depression. A later study by Weiss, Jackson, and Susser (1997) suggested that parents are less likely to report co-occurring internalizing and externalizing behaviors than would be expected given their rates of co-occurrence in the population, indicating that concern over one problem may decrease parents' awareness or concern over the other problem. Taken together, the findings of the two studies suggest that parents may be more concerned about externalizing than internalizing behaviors when their children exhibit both types of behavior. Given the traditional structure of a time-limited clinical interview, which typically begins with asking parents why they brought their children for treatment, it is possible that parents with children who have comorbid externalizing and internalizing problems may not spontaneously report (or emphasize) the internalizing problems because of their concern over the externalizing ones. In diagnostic interviews, then, internalizing symptoms may be less likely to be reported than externalizing ones, unless clinicians specifically ask about both types of symptoms. This could lead to better agreement for externalizing than for internalizing problems. This hypothesis was supported by the fact that for the clinician diagnoses, there was a negative relationship between internalizing disorders and externalizing disorders, indicating that clinicians were unlikely to diagnose these conditions together. No such relationship was found for co-occurring internalizing or co-occurring externalizing disorders, and the relationships between DISC diagnoses were nearly all positive, indicating that the DISC was quite likely to assign diagnoses together, regardless of whether they were internalizing or externalizing.

The preceding discussion has focused on factors external to the *DSM* system that may account for some of the diagnostic discordance in our data. None of these factors seem to adequately explain the massive level of disagreement we found between standardized interview diagnoses and clinician diagnoses. It may be, then, that we need to look to the system underlying these diagnoses for additional explanations. One potential problem is that many *DSM* diagnoses are based on lists of symptoms of which a certain number (e.g., five out of nine symptoms for major depressive episode) are required to meet criteria. This means that two people with rather different symptom presentations can be given the same diagnosis; the resulting within-diagnosis heterogeneity can add confusion to the diagnostic process (Clark, Watson, & Reynolds, 1995). The high rate of comorbidity among the disorders can also pose a problem (Clark et al., 1995); for example, the very low level of agreement we found for the depressive disorders could be related to the fact that depression in children rarely occurs alone (Hammen & Compas, 1994), particularly in clinic-referred children (Hammen, Rudolph, Weisz, Rao, & Burge, 1999). Given that clinicians are more often than not faced with a complex symptom presentation that crosses multiple diagnostic boundaries, a question could be raised as to how appropriate or useful it is to treat these disorders as distinct and go through the process of parsing out symptoms into their appropriate places within the *DSM* framework. Even if one decides that the task is appropriate, there can be little question that it is potentially difficult to carry out with precision, in the absence of a standardized interview procedure (e.g., in the conditions of usual clinical practice).

Although the present findings concern *DSM–III–R* diagnoses, there are reasons to suspect that the findings might not have differed substantially had other editions of the *DSM* been used, including *DSM–IV*. First, the basic concerns about the *DSM* noted in the previous paragraph (e.g., within-diagnosis heterogeneity and comorbidity between diagnoses) are relevant to other versions and were not resolved with the creation of *DSM–IV*. Second, studies of agreement across *DSM* versions have concluded that research findings using *DSM–III–R* are generalizable to *DSM–IV* diagnoses for the youth categories studied thus far (youth anxiety disorders; Kendall & Warman, 1996; ADHD; Biederman, Faraone, Weber, Rater, & Park, 1997). Third, the criteria for some categories (e.g., mood disorders) were essentially unchanged from *DSM–III–R* to *DSM–IV* (Callahan, Panichelli-Mindel, & Kendall, 1996). Thus, there are reasons to suspect that the low level of agreement we found with *DSM–III–R* would not improve with *DSM–IV*, but this should be tested directly in future research.

Given the importance of shared communication among and between researchers and clinicians, the present findings are a cause for concern. They suggest that the nature of the *DSM*, or the way it is being used, may lead to rather wide discrepancies between the diagnoses generated in regular clinical practice and those generated by systematic and standardized procedures designed to produce the most reliable assessment of child disorders. These discrepancies have several implications for the field of psychology. One implication is that some of the goals of the framers of the *DSM*—that is, (a) "clinical usefulness for making treatment and management decisions in varied clinical settings," (b) "reliability of the diagnostic categories" (American Psychiatric Association, 1987, p. xix), and (c) "enabl[ing] clinicians and investigators to. . . communicate about. . . the various mental disorders" (American Psychiatric Association, 1987, p. xxix)—have not yet been fully attained. While the portion of the *DSM* system that pertains to children is now sufficiently clear and explicit in its diagnostic

criteria to be used reliably across trained researcher interviewers (see, e.g., Shaffer et al., 1993), the present findings suggest that research interviewers and practicing clinicians may use the system in different ways and reach very different diagnostic conclusions about the same children.

An additional implication of these findings concerns dissemination of research to clinicians. If clinicians and researchers tend to assign different diagnoses to the same children, this may significantly undermine efforts to apply research findings to clients in practice. Clinical trials testing specific child treatments generally use samples of children assessed for diagnoses using standardized interview procedures. Clinicians who attempt to use the resulting treatments in their practice but who identify different children as treatment candidates (e.g., as having depressive disorders) than the clinical trials researchers would have identified, because they apply the *DSM* differently, risk applying the treatments to the wrong children. Similarly, practitioners who read descriptive psychopathology research to understand children in a particular diagnostic group may not find the research very helpful if the children they place in that diagnostic group do not actually correspond diagnostically to those studied by the researchers.

For many years, practitioners have expressed concern that clinical research has little relevance to actual clinical practice (see, e.g., Abrahamson, 1999; L. Cohen, 1979; Norcross, 1999). The present findings expand on this general concern, suggesting that one underlying problem may be a mismatch between diagnosis in practice and diagnosis in research. A useful objective for future inquiry will be to understand the mismatch and how to address it. Comparisons between children who receive a given diagnosis from either source alone or from both sources on variables such as comorbidity or dimensional measures of impairment could shed some light on the issue by illustrating the types of impairment clinicians and standardized interviews are indexing and where they differ. Analyses involving symptom patterns could clarify whether clinicians are differentially weighting various symptoms, unlike the DISC, which weights all symptoms equally, or assigning diagnoses at different symptom thresholds than the DISC. Studies such as these may help elucidate the nature of diagnosis in different settings and thus move our categorical system of diagnosis in a direction that will more fully meet the laudable goals of the *DSM* framers.

References

Abrahamson, D. J. (1999). Outcomes, guidelines, and manuals: On leading horses to water. *Clinical Psychology: Science and Practice, 6*, 467–471.

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. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.

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

Aronen, E. T., Noam, G. G., & Weinstein, S. R. (1993). Structured diagnostic interviews and clinicians' discharge diagnoses in hospitalized adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*, 674–681.

Biederman, J., Faraone, S. V., Weber, W., Rater, R. L., & Park, K. (1997). Correspondence between *DSM–III–R* and *DSM–IV* attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*, 1682–1687.

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

Boyle, M. H., & Pickles, A. R. (1997). Influence of maternal depressive symptoms on ratings of childhood behavior. *Journal of Abnormal Child Psychology, 25*, 399–412.

Breslau, N. (1987). Inquiring about the bizarre: False positives in Diagnostic Interview Schedule for Children (DISC) ascertainment of obsessions, compulsions, and psychotic symptoms. *Journal of the American Academy of Child and Adolescent Psychiatry, 26*, 639–644.

Callahan, S. A., Panichelli-Mindel, S. M., & Kendall, P. C. (1996). *DSM–IV* and internalizing disorders: Modifications, limitations, and utility. *School Psychology Review, 25*, 297–307.

Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., Crits-Christoph, P., et al. (1998). Update on empirically validated therapies II. *The Clinical Psychologist, 51*, 3–16.

Chilcoat, H. D., & Breslau, N. (1997). Does psychiatric history bias mothers' reports? An application of a new analytic approach. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*, 971–979.

Clark, L. A., Watson, D., & Reynolds, S. (1995). Diagnosis and classification of psychopathology: Challenges to the current system and future directions. *Annual Review of Psychology, 46*, 121–153.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cohen, L. (1979). The research readership and information source reliance of clinical psychologists. *Professional Psychology, 10*, 780–785.

Conrad, M., & Hammen, C. (1989). Role of maternal depression in perceptions of child maladjustment. *Journal of Consulting and Clinical Psychology, 57*, 663–667.

Costello, E. J., Edelbrock, C. S., & Costello, A. J. (1985). Validity of the NIMH Diagnostic Interview Schedule for Children: A comparison between psychiatric and pediatric referrals. *Journal of Abnormal Child Psychology, 13*, 579–595.

Derogatis, L. R. (1992). *The Brief Symptom Inventory: Administration, scoring and procedures manual—II*. Baltimore: Clinical Psychometric Research.

Edelbrock, C. S., Costello, A. J., Dulcan, M. K., Conover, N. C., & Kala, R. (1986). Parent–child agreement on child psychiatric symptoms assessed via structured interview. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 27*, 181–190.

Ezpeleta, L., De la Osa, N., Domenech, J. M., Navarro, J. B., Losillo, J. M., & Judez, J. (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 and Allied Disciplines, 38*, 431–440.

Fergusson, D. M., Lynskey, M. T., & Horwood, L. J. (1993). The effect of maternal depression on maternal ratings of child behavior. *Journal of Abnormal Child Psychology, 21*, 245–269.

Fisher, P. W., Shaffer, D., Piacentini, J., Lapkin, J., Kafantaris, V., Leonard, H., & Herzog, D. B. (1993). Sensitivity of the Diagnostic Interview Schedule for Children, 2nd edition (DISC–2.1) for specific diagnoses of children and adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*, 666–673.

Fleiss, J. R. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Hammen, C., & Compas, B. E. (1994). Unmasking unmasked depression in children and adolescents: The problem of comorbidity. *Clinical Psychology Review, 14*, 585–603.

Hammen, C., Rudolph, K., Weisz, J., Rao, U., & Burge, D. (1999). The context of depression in clinic-referred youth: Neglected areas in treatment. *Journal of the American Academy of Child and Adolescent Psychiatry, 38*, 64–71.

Herjanic, B., Herjanic, M., Brown, F., & Wheatt, T. (1975). Are children reliable reporters? *Journal of Abnormal Child Psychology, 3*, 41–48.

Herjanic, B., & Reich, W. (1982). Development of a structured psychiatric interview for children: Agreement between child and parent on individual symptoms. *Journal of Abnormal Child Psychology, 10*, 307–324.

Kendall, P. C., & Warman, M. J. (1996). Anxiety disorders in youth: Diagnostic consistency across *DSM–III–R* and *DSM–IV*. *Journal of Anxiety Disorders, 10*, 453–463.

Lonigan, C. J., Elbert, J. C., & Johnson, S. B. (1998). Empirically supported psychosocial interventions for children: An overview. *Journal of Clinical Child Psychology, 27*, 138–145.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*, 153–157.

Norcross, J. C. (1999). Collegially validated limitations of empirically validated treatments. *Clinical Psychology: Science and Practice, 6*, 472–475.

Pellegrino, J. F., Singh, N. N., & Carmanico, S. J. (1999). Concordance among three diagnostic procedures for identifying depression in children and adolescents with EBD. *Journal of Emotional and Behavioral Disorders, 7*, 118–127.

Piacentini, J., Shaffer, D., Fisher, P. W., Schwab-Stone, M., Davies, M., & Giola, P. (1993). The Diagnostic Interview Schedule For Children—Revised version (DISC—R): iii. Concurrent criterion validity. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*, 658–665.

Puig-Antich, J., & Chambers, W. (1978). *The Schedule for Affective Disorders and Schizophrenia for School-Age Children (Kiddie–SADS)*. New York: New York State Psychiatric Institute.

Rapee, R. M., Barrett, P. M., Dadds, M. R., & Evans, L. (1994). Reliability of the *DSM–III–R* childhood anxiety disorders using structured interview: Interrater and parent–child agreement. *Journal of the American Academy of Child and Adolescent Psychiatry, 33*, 984–992.

Reich, W., Shayka, J. J., & Taibleson, C. (1991). *Diagnostic Interview Schedule for Children—DICA–R*. Unpublished manuscript, Washington University, Division of Child Psychiatry.

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

Schwab-Stone, M. E., Fisher, P., Piacentini, J., Shaffer, D., Davies, M., & Briggs, M. (1993). Test–retest reliability (The Diagnostic Interview Schedule for Children—Revised Version, DISC–R): Part 2. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*, 651–657.

Schwab-Stone, M. E., Shaffer, D., Dulcan, M. K., Jensen, P. S., Fisher, P., Bird, H. R., Goodman, S. H., Leahy, 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.

Shaffer, D., Fisher, P. W., Piacentini, J., Schwab-Stone, M., & Wicks, J. (1989). *Diagnostic Interview Schedule for Children—Second revision (DISC–2)*. New York: New York State Psychiatric Institute.

Shaffer, D., Schwab-Stone, M., Fisher, P. W., Cohen, P., Piacentini, J., Davies, M., Conners, C. K., & Regier, D. (1993). The Diagnostic Interview Schedule for Children—Revised version (DISC–R): I. Preparation, field testing, interrater reliability, and acceptability. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*, 643–650.

Spitzer, R. L., Williams, J. B., Gibbon, M., & First, M. B. (1992). The Structured Clinical Interview for DSM–III–R (SCID): i. History, rationale, and description. *Archives of General Psychiatry, 49*, 624–629.

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry, 42*, 725–728.

Stavrakaki, C., Vargo, B., Roberts, N., & Boodoosingh, L. (1987). Concordance among sources of information for ratings of anxiety and depression in children. *Journal of the American Academy of Child and Adolescent Psychiatry, 26*, 733–737.

Strakowski, S. M., Hawkins, J. M., Keck, P. E. J., McElroy, S. L., West, S. A., Bourne, M. L., Sax, K. W., & Tugrul, K. C. (1997). The effects of race and information variance on disagreement between psychiatric emergency service and research diagnoses in first-episode psychosis. *Journal of Clinical Psychiatry, 58*, 457–463.

Verhulst, F. C., & Van der Ende, J. (1991). Assessment of child psychopathology: Relationships between different methods, different informants and clinical judgment of severity. *Acta Psychiatrica Scandinavica, 84*, 155–159.

Vitiello, B., Malone, R., Buschle, P. R., Delaney, M. A., & Behar, D. (1990). Reliability of *DSM–III* diagnoses of hospitalized children. *Hospital and Community Psychiatry, 41*, 63–67.

Weinstein, S. R., Stone, K., Noam, G. G., Grimes, K., & Schwab-Stone, M. (1989). Comparison of DISC with clinicians' *DSM–III* diagnoses in psychiatric inpatients. *Journal of the American Academy of Child and Adolescent Psychiatry, 28*, 53–60.

Weiss, B., Jackson, E. W., & Susser, K. (1997). Effect of co-occurrence on the referability of internalizing and externalizing problem behavior in adolescents. *Journal of Clinical Child Psychology, 26*, 198–204.

Weisz, J. R., & Weiss, B. (1991). Studying the "referability" of child clinical problems. *Journal of Consulting and Clinical Psychology, 59*, 266–273.

Welner, Z., Reich, W., Herjanic, B., Jung, K. G., & Amado, H. (1987). Reliability, validity, and parent–child agreement studies of the Diagnostic Interview for Children and Adolescents (DICA). *Journal of the American Academy of Child and Adolescent Psychiatry, 26*, 649–653.

Received October 2, 2000
Revision received February 10, 2001
Accepted May 14, 2001 ■