

Reliability of *DSM-IV* Anxiety and Mood Disorders: Implications for the Classification of Emotional Disorders

Timothy A. Brown
Boston University

Peter A. Di Nardo
State University of New York at Oneonta

Cassandra L. Lehman and Laura A. Campbell
Boston University

The reliability of current and lifetime *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; American Psychiatric Association, 1994) anxiety and mood disorders was examined in 362 outpatients who underwent 2 independent administrations of the Anxiety Disorders Interview Schedule for *DSM-IV*: Lifetime version (ADIS-IV-L). Good to excellent reliability was obtained for the majority of *DSM-IV* categories. For many disorders, a common source of unreliability was disagreements on whether constituent symptoms were sufficient in number, severity, or duration to meet *DSM-IV* diagnostic criteria. These analyses also highlighted potential boundary problems for some disorders (e.g., generalized anxiety disorder and major depressive disorder). Analyses of ADIS-IV-L clinical ratings (0–8 scales) indicated favorable interrater agreement for the dimensional features of *DSM-IV* anxiety and mood disorders. The findings are discussed in regard to their implications for the classification of emotional disorders.

Classification of emotional disorders has been an inexact science, reflected by the modest reliability of many diagnostic categories and marked changes in definitional criteria across editions of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*; American Psychiatric Association, 1987, 1994). The diagnostic criteria for all anxiety and mood disorders were revised to varying degrees in the current, fourth edition of the *DSM* (*DSM-IV*; American Psychiatric Association, 1994). Often, these revisions were guided by reliability findings from large-scale studies of disorders from the revised, third edition of the *DSM* (*DSM-III-R*; American Psychiatric Association, 1987; see Di Nardo, Moras, Barlow, Rapee, & Brown, 1993; Mannuzza et al., 1989; Williams et al., 1992). For example, in addition to the introduction of a formal typology of panic attacks (i.e., unexpected, situationally predisposed, situationally bound; cf. Barlow, Brown, & Craske, 1994), *DSM-IV* criteria for panic disorder and agoraphobia no longer include severity specifiers (i.e., mild, moderate, severe). This revision was based on findings that whereas generally good interrater consistency was noted for dimensional indica-

tors of panic frequency and agoraphobia severity (e.g., in Di Nardo et al., 1993, the correlation between independent dimensional ratings of agoraphobic avoidance was .81), application of these categorical severity specifiers was associated with considerable unreliability. For instance, in Di Nardo et al. (1993), higher reliability was observed for current *DSM-III-R* panic disorder collapsing across all levels of agoraphobic avoidance ($\kappa = .71$) than for each level of agoraphobia severity ($\kappa_s = .61, .70, .40$, for mild, moderate, and severe agoraphobia, respectively).

Indeed, it has been found that diagnostic unreliability of *DSM* disorders often does not stem from disagreement on the presence of defining symptoms but rather from difficulties applying categorical cutoffs to these inherently dimensional phenomena (e.g., *DSM* threshold for presence or absence of disorder based on sufficient distress or lifestyle impairment; application of *DSM* severity or course specifiers). In Di Nardo et al. (1993) many of the diagnostic disagreements involving social phobia and specific phobia were cases in which both interviewers noted clear features of these disorders but did not concur that these symptoms met the *DSM-III-R* interference or distress threshold (cf. Antony et al., 1994; Stein, Walker, & Forde, 1994).

Another important issue in the classification of emotional disorders is the diagnostic reliability of generalized anxiety disorder (GAD). In large scale studies entailing administration of two independent structured interviews, *DSM-III-R* GAD was associated with poor to fair reliability (kappas for current GAD were .27 in Mannuzza et al., 1989, .53 in Di Nardo et al., 1993, and .56 in Williams et al., 1992). These findings, along with data indicating that GAD has a comorbidity rate exceeding 80% (e.g., Brawman-Mintzer et al., 1993; Brown & Barlow, 1992), led to debate among researchers as to whether there was sufficient evidence of discrimi-

Timothy A. Brown, Cassandra L. Lehman, and Laura A. Campbell, Center for Anxiety and Related Disorders, Boston University; Peter A. Di Nardo, Department of Psychology, State University of New York at Oneonta.

We thank Bonnie Conklin, Patricia Miller, Jeanne Esler, Jonathan Lerner, Jessica Grisham, and David Barlow for their assistance with this study.

Correspondence concerning this article should be addressed to Timothy A. Brown, Center for Anxiety and Related Disorders, Boston University, 648 Beacon Street, 6th floor, Boston, Massachusetts 02215-2013. Electronic mail may be sent to tabrown@bu.edu.

nant validity to retain GAD as a diagnostic category in *DSM-IV* (Brown, Barlow, & Liebowitz, 1994). Although GAD remains a formal category in *DSM-IV*, its diagnostic criteria were revised substantially in an effort to define its boundary in relation to mood and adjustment disorders, anxiety disorders, and nonpathological worry. These revisions include the requirement that worry must be perceived by the person as uncontrollable (based on evidence that the parameter of uncontrollability distinguishes GAD worry from normal worry; Abel & Borkovec, 1995; Borkovec, 1994) and the reduction in the number of symptoms forming the associated symptom criterion from 18 to 6 (symptoms of autonomic arousal were eliminated [e.g., accelerated heart rate, shortness of breath]; symptoms of tension and negative affect were retained [e.g., muscle tension, feeling keyed up/on edge, irritability]). Although the decision to eliminate autonomic symptoms was data driven (Brown, Marten, & Barlow, 1995; Marten et al., 1993), researchers have raised concern that this revision may obfuscate the boundary between GAD and the mood disorders (Clark & Watson, 1991). This boundary issue is reflected in a *DSM-IV* exclusionary criterion stating that GAD should not be assigned if its features occur exclusively during the course of a mood disorder. Nonetheless, it is important to determine whether the substantial changes to GAD in *DSM-IV* have resulted in improved diagnostic reliability.

Similarly, it would be of interest to evaluate what impact other modifications to the diagnostic definitions of emotional disorders have had on their reliability. Although the category of specific phobia has historically been associated with favorable interrater agreement (e.g., $\kappa > .80$ in Di Nardo et al., 1993, and Mannuzza et al., 1989), *DSM-IV* now requires that this diagnosis be assigned as one of the following types: (a) animal (e.g., dogs, rats); (b) natural environment (e.g., heights, storms); (c) blood/injury/injection (e.g., having a blood test); (d) situational (e.g., driving, enclosed places); and other (e.g., illness, vomiting). Although specification of specific phobia types was intended to account for the heterogeneity of the disorder, research is needed on the reliability and validity of these distinctions (cf. Antony, Brown, & Barlow, 1997).

The purpose of this study was to evaluate the reliability and factors contributing to diagnostic disagreements of the *DSM-IV* anxiety and mood disorders using the Anxiety Disorders Interview Schedule for *DSM-IV*: Lifetime version (ADIS-IV-L; Di Nardo, Brown, & Barlow, 1994). The revisions in the ADIS-IV-L go well beyond updating the Anxiety Disorders Interview Schedule—Revised (ADIS-R; Di Nardo & Barlow, 1988) to be consistent with *DSM-IV* criteria. Unlike the ADIS-R, the ADIS-IV-L provides diagnostic assessment of a broader range of conditions (e.g., substance use disorders), evaluation of lifetime disorders, and dimensional assessment of the key and associated features of disorders, irrespective of whether a formal *DSM-IV* diagnosis is under consideration (see Method section). The latter revision is based on the position that many features of emotional disorders operate on a continuum rather than in a categorical, presence/absence fashion as in *DSM* diagnosis (cf. Brown, 1996; Brown, Chorpita, & Barlow, 1998; Costello, 1992). Because of the importance of these dimensional ratings as indicators in clinical trials and nosology and psychopathology studies (e.g., Borkovec & Costello, 1993; Brown et al., 1998), another aim of this study was to examine the interrater reliability of these measures.

Method

Participants

Participants were 362 patients presenting for assessment and treatment at the Center for Stress and Anxiety Disorders, University at Albany, State University of New York ($n = 70$), and the Center for Anxiety and Related Disorders, Boston University ($n = 292$).¹ (The two research centers are collectively referred to as “the center.”) Women constituted the larger portion of the sample (58%); average age was 33.11 ($SD = 10.62$, range = 18 to 62). The racial and ethnic breakdown of the sample was Caucasian (88%), African American (4%), Hispanic (3%), Asian (3%), Pacific Islander (1%), and other or missing (2%).

Patients were required to meet several inclusion and exclusion criteria that were assessed by telephone screening at initial contact with the center and reassessed and confirmed during the diagnostic interviews. Specifically, patients were required to be between the ages of 18 and 65 and to have a presenting complaint that likely involved an anxiety or mood disorder. Patients were excluded from the study if any of the following were present: (a) current hallucinations or delusions, (b) current or recent (within the past 6 months) alcohol or substance abuse or dependence, (c) current suicidal or homicidal risk meriting crisis intervention, and (d) two or more hospitalizations in the past 5 years for psychotic symptoms. Patients were also required to meet psychotropic medication and psychotherapy stabilization criteria for the periods preceding and overlapping with the diagnostic assessment. Patients using anxiolytics and beta-blockers were required to maintain the same dosage for at least 1 month. Patients on antidepressants (tricyclics, selective serotonin reuptake inhibitors, and monoamine oxidase inhibitors) had to maintain a stable dosage for at least 3 months. The medication wash out period (i.e., period since medication discontinuation) was 1 month for all medications. Patients in psychotherapy for an emotional problem were required to satisfy a 3-month stabilization period; the psychotherapy wash out period was 1 month.

The current sample was randomly selected to receive two independent ADIS-IV-L interviews from roughly 1,400 consecutive admissions to the center who met eligibility criteria between the periods of December 1994 and October 1999. In most cases (79%), the second ADIS-IV-L occurred within 2 weeks of the first interview ($M = 10.60$ days, $SD = 8.60$). After both interviews had been completed and the interviewers had independently recorded their diagnostic judgments, cases were presented in weekly staff meetings that entailed the presentation of interviewers' diagnoses, discussion of factors contributing to any diagnostic disagreements, and establishment of consensus diagnoses. The primary source of unreliability for each diagnostic disagreement was recorded (by Timothy A. Brown or Peter A. Di Nardo) using a rating system designed for use in the present study: (a) difference in report—patient gives different information to the two interviewers (e.g., variability in responses to inquiry about the presence, severity, or duration of key symptoms); (b) threshold—consistent symptom report is provided across interviews, but interviewers disagree on whether these symptoms cause sufficient interference and distress to satisfy the *DSM-IV* threshold for a clinical disorder; (c) change in clinical status—clear change in the severity or presence of symptoms between interviews; (d) interviewer error—interviewer improperly applies *DSM-IV* diagnostic or exclusion rules or fails to obtain necessary diagnostic information during ADIS-IV-L administration (e.g., skips an ADIS-IV-L diagnostic section prematurely); (e) diagnosis subsumed under another condition—disagreement on whether symptoms are attributable to, or better accounted for by, a co-occurring disorder; and (f) *DSM-IV* in clarity—disagreement stems from limitations of the *DSM-IV* criteria in providing clear direction for differential diagnosis.

¹ Our research center relocated from the University at Albany, State University of New York, to Boston University in September 1996.

Anxiety Disorder Interview Schedule for DSM-IV: Lifetime Version (ADIS-IV-L; Di Nardo et al., 1994)

The ADIS-IV-L is a semistructured interview designed to establish reliable diagnosis of the *DSM-IV* anxiety, mood, somatoform, and substance use disorders and to screen for the presence of other conditions (e.g., psychotic disorders). The ADIS-IV-L is a substantial revision of the ADIS-R. In addition to being updated for *DSM-IV* criteria, the ADIS-IV-L provides assessment of lifetime disorders and a diagnostic timeline that fosters accurate determination of the onset, remission, and temporal sequence of current and lifetime disorders. Moreover, in several ADIS-IV-L sections, raters make dimensional ratings (0–8) of disorder features regardless of whether a *DSM-IV* diagnosis is under consideration. This occurs in the following sections: (a) social phobia—ratings of fear or avoidance of 13 social situations; (b) generalized anxiety disorder—ratings of excessiveness and difficulty controlling worry in 8 areas; (c) obsessive-compulsive disorder—ratings of persistence, distress, and resistance of 9 obsession types and frequency of 6 compulsions; and (d) specific phobia—ratings of fear or avoidance of 17 objects or situations from the 5 types of *DSM-IV* specific phobias (animals, natural environment, blood/injection/injury, situational, other).

Dimensional ratings of the features of panic disorder and agoraphobia are completed by interviewers only if these diagnoses are under consideration (otherwise, the interviewer would skip this diagnostic section after receiving negative responses to initial screening questions). Ratings in the panic disorder and agoraphobia sections include (a) frequency of panic attacks in the past month, (b) fear of panic attacks in the past month (0–8 scale), and (c) current avoidance of or escape from 22 agoraphobic situations (0–8 scale). Dimensional ratings (0–8 scales) in the major depression and dysthymia sections and the associated symptoms portion of the generalized anxiety disorder section are arranged in the same fashion as the panic disorder and agoraphobia sections of the ADIS-IV-L. However, for purposes of the present and other ongoing studies, in the Boston University sample ($n = 292$), interviewers inquired about and assigned these ratings regardless of whether a mood or generalized anxiety disorder diagnosis was under consideration. These ratings were as follows: (a) major depression—ratings of the seven symptoms that accompany depressed mood and diminished interest and pleasure in activities to form the key criterion of major depressive episode; (b) dysthymia—ratings of the six symptoms comprising its associated symptom criterion; and (c) generalized anxiety disorder—ratings of the frequency and severity of the six symptoms comprising its associated symptoms criterion. In these and other ADIS-IV-L sections, interviewers followed the appropriate *DSM-IV* duration criterion (e.g., more days than not for a period of 2 years or greater in dysthymia) when making dimensional ratings (i.e., ratings reflected a composite of severity, frequency, or duration in respect to the *DSM-IV* criterion, if specified).

For each current and lifetime diagnosis, interviewers assigned a 0–8 clinical severity rating that indicated their judgment of the degree of distress and interference in functioning associated with the disorder (0 = none to 8 = very severely disturbing/disabling). In instances in which the patient met criteria for two or more current diagnoses, the principal diagnosis was the one that received the highest clinical severity rating. For both current and lifetime disorders, those that met *DSM-IV* criteria for a formal diagnosis were assigned clinical severity ratings of 4 (definitely disturbing/disabling) or higher (clinical diagnoses). Current clinical diagnoses that were not deemed to be the principal diagnosis are referred to as additional diagnoses. When the key features of a current or lifetime disorder were present but were not judged to be extensive or severe enough to warrant a formal *DSM-IV* diagnosis (or for *DSM-IV* disorders in partial remission), clinical severity ratings of 1–3 were assigned (subclinical diagnoses). When no features of a disorder were present, clinical severity ratings of 0 were given.

Interviewers

Diagnosticians were 6 doctoral-level clinical psychologists and 30 advanced clinical doctoral students. Before participating in the study, diagnosticians were required to undergo extensive training and meet strict certification criteria in the administration of the ADIS-IV-L. Training began with the trainees reading the ADIS-IV-L manual, observing videotaped interviews, and then observing at least three live ADIS-IV-L interviews conducted by a senior, certified interviewer. While observing live interviews, the trainee made ratings and diagnoses. After the interview, the trainee and senior interviewer compared and discussed diagnoses and dimensional ratings. Following observation of several live interviews, trainees had the option to administer one or more collaborative interviews to become more comfortable with ADIS-IV-L administration prior to the certification phase. In a collaborative interview, the trainee assumed primary responsibility for ADIS-IV-L administration, but the senior interviewer could interject as needed (e.g., ask differential diagnosis questions the trainee had not asked or provide an indication of when to skip a diagnostic section). In the certification phase, trainees were required to administer a minimum of three ADIS-IV-Ls under observation of a senior interviewer. After the interview, the trainee and senior interviewer independently established current and lifetime diagnoses.

The criteria for ADIS-IV-L certification was that within three of five consecutive interviews, the trainee's diagnoses must match the senior interviewers' diagnoses and the trainee must commit no ADIS-IV-L administration errors based on a checklist of nine items (e.g., omission of mandatory inquiry or failure to ask necessary follow-up questions of clarification). A match was defined as (a) agreement on the principal diagnosis (including *DSM-IV* severity descriptors such as major depression, single episode, moderate) and agreement within 1 point on its clinical severity rating and (b) identification as a clinical disorder all additional and lifetime diagnoses assigned by the senior interviewer as meeting the *DSM-IV* threshold (i.e., clinical severity rating ≥ 4). Agreement on the clinical severity ratings of additional and lifetime diagnoses was not required, and the trainee was not required to match with the interviewer on diagnoses not formally assessed by the ADIS-IV-L (e.g., sexual disorders, eating disorders). Interviews were classified as failing toward certification when the trainee was rated as having committed one or more administration errors, regardless of whether his or her diagnoses matched those of the senior interviewer.

Results

Reliability of DSM-IV Diagnostic Categories

Current diagnoses. Interrater reliability of *DSM-IV* diagnoses was calculated by kappa coefficients using the formula presented in Fleiss, Nee, and Landis (1979). Following the guidelines used in studies of the reliability of *DSM-III-R* anxiety and mood disorders (e.g., Di Nardo et al., 1993; Mannuzza et al., 1989), the standards used to interpret kappa coefficients were as follows: excellent agreement ($\kappa \geq .75$), good agreement ($.60 \leq \kappa \leq .74$), fair agreement ($.40 \leq \kappa \leq .59$), and poor agreement ($\kappa < .40$).

In Table 1 we present reliability findings for current *DSM-IV* diagnoses. For purposes of comparison, the findings from our reliability study of *DSM-III-R* anxiety and mood disorders (Di Nardo et al., 1993) are also provided in Table 1. Using the aforementioned standards, we found that all principal diagnoses evidenced good or excellent reliability with the exception of dysthymia (DYS), although the kappas for panic disorder (PD) and DYS should be interpreted cautiously because these categories were assigned infrequently as principal diagnoses in the sample ($n_s = 14$ and 15, respectively).

Table 1
Diagnostic Reliability of Current DSM-IV Diagnoses (N = 362)
and Current DSM-III-R Diagnoses (N = 267)

Diagnostic category	Principal diagnosis				Principal or additional diagnosis			
	DSM-IV		DSM-III-R ^a		DSM-IV		DSM-III-R ^a	
	κ	<i>n</i>	κ	<i>n</i>	κ	<i>n</i>	κ	<i>n</i>
PD	.72	14	.43	38	.56	22	.39	44
PDA	.77	83	.72	131	.81	102	.71	142
PD & PDA	.79	94	.79	152	.79	120	.75	168
Specific phobia	.86	56	.82	21	.71	100	.63	47
Social phobia	.77	80	.79	45	.77	152	.66	84
GAD	.67	76	.57	38	.65	113	.53	108
OCD	.85	33	.80	19	.75	60	.75	24
PTSD	—	—	.46	3	.59	14	.55	8
MDD	.67	53	.65	8	.59	111	.55	46
DYS	.22	15	-.05	5	.31	53	.35	25
MDD & DYS	.72	61	.46	13	.63	138	.56	64

Note. *n* = number of cases in which diagnosis was assigned by either or both raters; dashes indicate an insufficient *n* to calculate kappa; PD = panic disorder; PDA = panic disorder with agoraphobia; GAD = generalized anxiety disorder; OCD = obsessive-compulsive disorder; PTSD = posttraumatic stress disorder; MDD = major depressive disorder; DYS = dysthymia.

^aData are from Di Nardo, Moras, Barlow, Rapee, and Brown (1993).

With the exception of social phobia (SOC), which continued to be associated with excellent interrater agreement, higher kappas were observed for all principal DSM-IV anxiety and mood disorders relative to reliability findings for the corresponding DSM-III-R categories. The most substantial improvement (i.e., from fair to good reliability) was evident for the principal diagnoses of PD (from .43 in DSM-III-R to .72 in DSM-IV), generalized anxiety disorder (GAD; from .57 in DSM-III-R to .67 in DSM-IV); and mood disorders (collapsing major depressive disorder [MDD] and DYS; from .46 in DSM-III-R to .72 in DSM-IV). However, *z* tests of the differential magnitude of these kappas did not reach statistical significance.

We noted a similar pattern of results when examining any current clinical disorder, collapsing across principal and additional diagnoses (see Table 1). Excellent reliability was obtained for panic disorder with agoraphobia (PDA), obsessive-compulsive disorder (OCD), SOC, and panic disorder collapsing across the presence or absence of agoraphobia (PD and PDA). The categories associated with good reliability were specific phobia (SPEC), GAD, and any mood disorder (MDD and DYS). Fair reliability was found for PD, MDD, and posttraumatic stress disorder (PTSD); DYS continued to be associated with poor reliability. As was the case for principal diagnoses only, we obtained higher kappas (albeit not statistically significant as evaluated by *z* tests) for all DSM-IV categories relative to DSM-III-R, with the exception of DYS which went from .35 to .31 and OCD, which did not change ($\kappa = .75$ in both studies).

Disorder types and specifiers. Most DSM-IV categories include additional subclassifications to indicate the nature, course, or severity of the disorder. The reliability of these subtypes and specifiers was examined for any current clinical disorder (i.e., principal or additional diagnosis). We evaluated the interrater agreement of the specific phobia types and the generalized type of social phobia using the entire sample. For MDD and DYS, reli-

ability of specifiers was examined in cases in which both interviewers assigned the disorder at a clinical or subclinical level (i.e., specifiers are only recorded when MDD or DYS is diagnosed).

The results of these analyses are presented in Table 2. At the level of principal diagnosis, excellent reliability was obtained for

Table 2
Diagnostic Reliability of Current DSM-IV
Diagnostic Types and Specifiers

Type or specifier	Principal diagnosis		Principal or additional diagnosis	
	κ	<i>n</i> ^a	κ	<i>n</i> ^a
Specific phobia				
Animal	.80	3	.53	16
Natural environment	.85	8	.53	21
Blood, injury, or injection	1.00	4	.66	10
Situational	.86	31	.73	58
Other	.89	10	.96	13
Social phobia				
Generalized	.73	43	.73	89
Major depressive disorder				
Single or recurrent	.46	26	.55	55
Mild, moderate, or severe	.30	26	.36	55
Chronic or nonchronic	.62	26	.67	55
Dysthymia				
Early or late onset	—	—	.55	9

Note. Dashes indicate an insufficient *n* to calculate kappa.

^aFor analyses of specific phobia and social phobia types, *n* refers to the number of cases in which the type was assigned by either or both raters in the total study sample (*N* = 362); for major depressive disorder and dysthymia, *n* refers to size of the subsample (i.e., number of patients assigned the disorder by both raters at the clinical or subclinical level) used in the analysis of specifiers.

each of the specific phobia types, although these findings should be interpreted with caution given the small sample sizes associated with some analyses. For all but the other type, these estimates decreased when reliability was examined using any current clinical disorder. Consistent with a previous finding using *DSM-III-R* definitions ($\kappa = .69$; Mannuzza et al., 1995), the generalized type of *DSM-IV* social phobia evidenced good reliability both as a principal diagnosis and as a current diagnosis at any clinical level ($\kappa_s = .73$).

Interrater agreement for the course and severity specifiers of MDD and DYS is also presented in Table 2. Whereas the course and onset specifiers for MDD and DYS were associated with fair to good reliability (range of $\kappa_s = .46$ to $.67$), poor reliability was found for the MDD severity specifier ($\kappa_s = .30$ and $.36$). Reliability for the early/late onset specifier of principal DYS and specifiers for other disorders (e.g., poor insight in OCD) could not be estimated because of the excessively low rate that either the diagnosis or specifier was assigned in the sample.

Sources of unreliability. Factors contributing to diagnostic disagreements were evaluated in current clinical diagnoses (collapsing principal and additional status). As can be seen in Table 3, the prevailing sources of unreliability differed substantially across the anxiety and mood disorders. For instance, the majority of disagreements involving SOC, SPEC, and OCD (62% to 67%) entailed cases in which one interviewer assigned the diagnosis at a clinical level and the other rated the diagnosis as subclinical; for other categories (e.g., PDA, GAD, MDD, DYS), this was a relatively rare source of unreliability. Indeed, the "threshold" issue was the most common source of disagreements for the diagnoses of SPEC and SOC. Difference in patient report was otherwise the most prevalent source of unreliability, ranging from 22% in SPEC to 100% in PTSD. Differential aggregation of unreliability sources was found for change in clinical status as well; although a rare source for other disorders, it accounted for 9 of the 53 (17%) MDD disagreements, consistent with the episodic nature of this condition.

Considerable variability was also evident across categories for the frequency with which other disorders were involved in diagnostic disagreements. Whereas disagreements with other disorders were relatively uncommon for SOC, OCD, and PTSD (8% to 13%), another clinical diagnosis was involved in over half of the disagreements with DYS, PDA, MDD, and GAD (54% to 74%). Table 3 provides the specific disorders that were involved in these disagreements for each diagnosis. As can be seen in this table, disagreements entailing another clinical diagnosis quite often involved disorders that had overlapping definitional features and that differed mainly in the duration or severity of symptoms (e.g., PD vs. PDA; SPEC vs. agoraphobia without a history of PD; MDD vs. DYS). In addition, this overlap was evident in disagreements involving anxiety disorder not otherwise specified (NOS) and depressive disorder NOS diagnoses. For example, a category frequently involved in disagreements with GAD was anxiety disorder NOS (GAD; $n = 10$), in which one interviewer noted clinically significant features of GAD (i.e., clinical severity rating ≥ 4) but judged that not all criteria for a formal *DSM-IV* GAD diagnosis had been met (e.g., number or duration of worries or associated symptoms). This was also the case for the NOS diagnoses associated with disagreements in other disorders (e.g., in the two OCD disagreements involving another disorder, both were with anxiety disorder NOS [OCD]).²

Consistent with prior evidence that mood disorders may pose the greatest boundary problem for GAD, 22 of the 35 GAD

disagreements (63%) involving another diagnosis were with mood disorders (DYS = 10, MDD = 9, depressive disorder NOS = 2, bipolar = 1). Conversely, although most MDD disagreements involved other diagnoses (34 of 53), rarely ($n = 3$) were these disagreements with anxiety disorders. Indeed, most MDD disagreements involved other mood disorders (depressive disorder NOS = 15, DYS = 12). As shown in Table 3, other mood disorders were the most frequent diagnoses involved in DYS disagreements as well, although disagreements with GAD were more common ($n = 6$).

Lifetime diagnoses. In Table 4 we present findings for the reliability of lifetime diagnoses (i.e., collapsing across current and past diagnoses). Because alcohol and substance use disorders were assigned frequently as past diagnoses, it was possible to evaluate the reliability of these categories (the reliability of current alcohol and substance use disorders could not be examined because of a study exclusion criterion). Excellent reliability was obtained for PDA, panic disorder collapsing across the presence or absence of agoraphobia (PD and PDA), OCD, alcohol abuse or dependence, and substance abuse or dependence. SPEC, SOC, GAD, PTSD, MDD, and any mood disorder (MDD and DYS) were associated with good reliability. Fair reliability was found for PD. The lifetime diagnosis of DYS evidenced poor interrater agreement.

Reliability and Structure of DSM-IV Dimensional Features

Data reduction and factor analysis. We examined the interrater reliability of the dimensional ratings of *DSM-IV* anxiety and mood disorders features using the Boston University sample ($n = 292$). Prior to conducting reliability analyses, the ratings from each ADIS-IV-L section were submitted to factor analysis to provide an empirical basis for the formation of composite scores (principal-components extraction with oblique rotation, when needed).³ In most instances, unidimensional solutions were obtained, and these ADIS-IV-L sections were scored accordingly. However, analyses of OCD and SPEC ratings produced multifactorial structures. Consistent with prior evidence of the multidimensionality of these symptoms (e.g., Summerfeldt, Richter, Antony, & Swinson, 1999), a three-factor solution was obtained for persistence and distress ratings of the nine types of OCD obsessions (three items each): (a) contamination, doubting, accidental harm to others; (b) aggressive and nonsensical impulses and sexual

² To foster the descriptiveness of the anxiety disorder NOS and depressive disorder NOS categories, a diagnostic convention in our center is to specify (in parentheses) the formal *DSM-IV* category to which the NOS diagnosis is closest; for example, depressive disorder NOS (DYS) would be assigned in a case in which clinically significant features of DYS are present (i.e., clinical severity rating ≥ 4) but one or more of the *DSM-IV* criteria for DYS are not met (e.g., duration of slightly less than 2 years).

³ Analyses were limited to the Boston University subsample to ensure complete data for all contributing cases (i.e., ratings from the MDD, DYS, and associated symptoms of GAD sections were collected on a listwise basis in the Boston University sample only; see Method section). For the sake of brevity, details on the conduct and results of these factor analyses have been omitted from this report. A full description of factor analytic results and a comprehensive list of ADIS-IV-L ratings are available by written request to Timothy A. Brown.

Table 3
Factors Contributing to Diagnostic Disagreements for Current DSM-IV Anxiety and Mood Disorders

Disorder or source	Diagnosis																	
	PD		PDA		SPEC		SOC		GAD		OCD		PTSD		MDD		DYS	
	Pro.	n	Pro.	n	Pro.	n	Pro.	n	Pro.	n	Pro.	n	Pro.	n	Pro.	n	Pro.	n
Disagreements involving another clinical diagnosis																		
PD	.31	4	.15	4	—	—	—	—	.02	1	—	—	—	—	—	—	—	—
PDA	—	—	—	—	.08	3	—	—	—	—	—	—	—	—	—	—	—	—
SPEC	—	—	.19	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—
SOC	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
GAD	—	—	.04	1	.03	1	—	—	—	—	—	—	—	.04	2	.15	6	—
OCD	—	—	.04	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—
PTSD	—	—	—	—	—	—	—	—	.02	1	—	—	—	—	—	—	.39	16
MDD	—	—	—	—	—	—	—	—	.19	9	—	—	—	.23	12	—	—	—
DYS	—	—	—	—	—	—	—	—	.21	10	—	—	—	.28	15	.12	5	—
Depression NOS	—	—	—	—	—	—	—	—	.04	2	—	—	—	.02	1	—	—	—
Anxiety NOS	.08	1	.08	2	—	—	.05	2	.21	10	.10	2	.13	1	—	—	—	—
AGw/oPD	—	—	.04	1	.14	5	.03	1	—	—	—	—	—	—	—	—	—	—
Adjustment disorder	—	—	—	—	—	—	—	—	.02	1	—	—	—	.06	3	.02	1	—
Other	—	—	—	—	—	—	—	—	BIP	—	—	—	—	PAIN	—	SOM	—	—
Total	.38	5	.54	14	.27	10	.08	3	.74	35	.10	2	.13	1	.64	34	.71	29
Clinical vs. subclinical disagreements ^a	.46	6	.12	3	.62	23	.67	26	.13	6	.62	13	.25	2	.17	9	—	—
Sources of unreliability																		
Threshold ^b	.46	6	.12	3	.46	17	.41	16	.09	4	.29	6	—	—	.15	8	.05	2 ^d
Difference in patient report	.31	4	.50	13	.22	8	.36	14	.55	26	.48	10	1.00	8	.55	29	.66	27
Interviewer error	.23	3	.19	5	.14	5	.10	4	.19	9	.14	3	—	—	.11	6	.15	6
Diagnosis subsumed under comorbid disorder	—	—	.15	4	.08	3	.10	4	.15	7	.10	2	—	—	—	—	.15	6
Change in clinical status	—	—	.04	1	.03	1	—	—	.02	1	—	—	—	—	.17	9	—	—
DSM-IV in clarity	—	—	—	—	.05	2	.03	1	—	—	—	—	—	—	—	—	—	—
Missing	—	—	—	—	.03	1	—	—	—	—	—	—	—	—	—	—	—	—
Total disagreements	—	13	—	26	.03	37	—	39	—	47	—	21	—	8	.02	1	—	41

Note. pro. = proportion; PD = panic disorder; PDA = panic disorder with agoraphobia; SPEC = specific phobia; SOC = social phobia; GAD = generalized anxiety disorder; OCD = obsessive-compulsive disorder; PTSD = posttraumatic stress disorder; MDD = major depressive disorder; DYS = dysthymia; Depression NOS = depressive disorder not otherwise specified; Anxiety NOS = anxiety disorder not otherwise specified; AGw/oPD = agoraphobia without a history of panic disorder; HYPO = hypochondriasis; BIP = bipolar disorder; PAIN = pain disorder; SOM = somatization disorder. Dashes indicate that the unreliability source was not present.

^a Cases in which both raters recorded the diagnosis but disagreed on whether it should be assigned at the clinical level. ^b Note that for SPEC and other diagnoses, the frequency of cases in which "Threshold" was the primary source of unreliability does not necessarily equal the number of disagreements entailing assignment of the disorder at clinical and subclinical levels. This is because each of the other unreliability sources (e.g., change in clinical status, variability in patient report regarding number or severity of symptoms) could also result in clinical versus subclinical disagreements. ^c This disagreement pertained to a boundary issue between specific phobia, other type (contracting an illness), and hypochondriasis. ^d In these 2 cases, disagreements stemmed from a threshold issue of MDD (chronic) versus DYS; that is, whether the features were sufficiently severe to be classified as a chronic major depressive episode.

Table 4
Diagnostic Reliability of Lifetime DSM-IV Diagnoses (N = 362)

Lifetime diagnosis	κ	<i>n</i>
Panic disorder (PD)	.58	30
Panic disorder with agoraphobia (PDA)	.81	116
PD & PDA	.79	120
Specific phobia	.70	114
Social phobia	.73	161
Generalized anxiety disorder	.65	114
Obsessive-compulsive disorder	.75	73
Posttraumatic stress disorder	.61	26
Major depressive disorder (MDD)	.68	208
Dysthymia (DYS)	.36	66
MDD & DYS	.69	224
Alcohol abuse or dependence	.83	47
Substance abuse or dependence	.82	48

Note. *n* = number of cases in which diagnosis was assigned by either or both raters.

thoughts or impulses; and (c) nonsensical thoughts/images, horrific images, and religious/Satanic thoughts/impulses. A two-factor solution was obtained for the frequency ratings of six OCD compulsions. This structure entailed (a) the five compulsions of checking, washing, adhering to rules or sequences, internal repetition, and counting; and (b) the single compulsion of hoarding (cf. Baer, 1994).

A four-factor solution was obtained for the fear ratings of 17 SPEC objects and situations: (a) blood/injury/injection (6 items: blood from cut, receiving injections, having blood drawn—either in self or others); (b) situational (5 items: elevators/enclosed places, air travel, driving, storms, heights); (c) Illness (3 items: vomiting, contracting an illness, choking); and (d) animals or water (2 items). However, because animal fears and water fears were quite modestly correlated ($r = .15$) and because there was not a clear conceptual basis for collapsing these ratings, they were evaluated separately in reliability analyses. In addition, fear of dental or medical procedures did not have a salient loading on any factor and was thus analyzed separately.

Interrater reliability of ADIS-IV-L dimensional ratings. In Table 5 we provide reliability estimates (Pearson r s) for dimensional ratings of DSM-IV anxiety and mood disorder features. Although all are included for informational purposes, the composite scores that pertained to different parameters of the same items were highly overlapping. Specifically, the following intercorrelations were noted: (a) social phobia fear versus avoidance ratings ($r = .95$), (b) specific phobia fear versus avoidance ratings (range of r s = .76 to .90), (c) excessiveness versus uncontrollability of GAD worry ($r = .91$), and (d) persistence or distress versus resistance of OCD obsessions (range of r s = .85 to .94).

Acceptable interrater reliability was found for the majority of the various dimensional ratings. In most cases, the lowest estimates were for single-item ratings such as specific phobia avoidance of dental or medical procedures (.41) and avoidance of water (.48). The findings from reliability analyses of the 9-point (0–8) ADIS-IV-L clinical severity rating for each disorder are also shown in Table 5. Quite favorable reliability was obtained for the clinical severity ratings of most disorders. However, consistent with findings at the diagnostic level, reliability of the DYS clinical severity rating was low ($r = .36$).

Discussion

Diagnostic Reliability of Current and Lifetime DSM-IV Anxiety and Mood Disorders

Collectively, these findings suggest that most current disorders are associated with good to excellent interrater agreement.⁴ For example, all principal diagnostic categories except DYS evidenced good to excellent reliability. In comparison with our DSM-III-R reliability study (Di Nardo et al., 1993), improved reliability was noted for the vast majority of DSM-IV disorders, and no DSM-IV category was associated with a markedly lower reliability estimate. Diagnoses showing the most improved reliability were PD and GAD. As was the case for current diagnoses, good to excellent reliability was found for the majority of lifetime anxiety and mood disorders. Interestingly, excellent interrater agreement was obtained for the alcohol and substance use disorders (κ s = .83 and .82, respectively), indicating the potential utility of the ADIS-IV-L to provide reliable DSM-IV diagnosis of these conditions.

The improved reliability of GAD is particularly encouraging because this category was in jeopardy of being removed from DSM-IV, in part because of the poor to fair reliability of its DSM-III-R definition. This improvement could perhaps be attributed to the revised definition of GAD in DSM-IV, which emphasizes the uncontrollable nature of worry and the associated symptoms of tension and negative affect. However, GAD diagnostic disagreements frequently involved the mood disorders (47%).⁵ This is consistent with prior evidence (e.g., Brown et al., 1998; Starcevic, 1995) that the mood disorders pose a more significant boundary issue for GAD than do other anxiety disorders. In future research, it would be important to examine the discriminant validity of GAD and mood disorders and determine if the diagnostic definition of GAD could be further refined to foster its distinction from these conditions. Also noteworthy is the finding that difference in patient report was rated the most common source of GAD disagreement (55%). This finding could also be reflective of limitations in the diagnostic criteria. Reliable diagnosis of GAD requires consistent self-report of many subjective features (e.g., number and severity of worry areas and physical symptoms) and their onset and duration in relation to other conditions (e.g., mood disorders). Inconsistency in such reports could be indicative of vagueness of these diagnostic features and patients' difficulty differentiating them from other disorders. Bearing on this point, previous research has shown that disorders associated with clear behavioral markers (e.g., OCD with compulsions, and situational avoidance in PDA, SOC, or SPEC) are associated with higher reliability than disorders without such features (e.g., PD, GAD,

⁴ One could argue that the present rates of interrater agreement represent the upper limit of potential reliability estimates for these disorders given aspects of the study methodology such as use of highly trained interviewers and the specialized anxiety and mood disorders setting (i.e., diagnostic reliability might be lower in primary clinical settings that often entail patient populations of a wider range of disorders, less structured clinical assessments, etc.).

⁵ It is noteworthy that none of the GAD disagreements involved OCD (or vice versa) despite previous concerns about boundary problems with excessive worry and obsessions (Brown, Moras, Zinbarg, & Barlow, 1993; Turner, Beidel, & Stanley, 1992).

Table 5
Interrater Reliability of ADIS-IV-L Dimensional Ratings of DSM-IV Disorder Features

Feature or rating	<i>r</i>
Panic disorder/agoraphobia	
Number of panic attacks (past month)	.58
Fear of panic attacks (past month)	.53
Agoraphobic avoidance	.86
Clinical severity rating	.83
Social phobia	
Situational fear	.86
Situational avoidance	.86
Clinical severity rating	.80
Generalized anxiety disorder	
Excessive worry	.73
Uncontrollability of worry	.78
Associated symptoms	.83
Clinical severity rating	.72
Obsessive-compulsive disorder	
Obsessions: persistence distress	
Doubting, contamination, accidental harm	.75
Impulses (aggressive, sexual, nonsensical)	.68
Other (religious, horrific, nonsensical thoughts)	.78
Obsessions: resistance	
Doubting, contamination, accidental harm	.76
Impulses (aggressive, sexual, nonsensical)	.43
Other (religious, horrific, nonsensical thoughts)	.72
Compulsions	
Compulsion frequency	.79
Hoarding frequency	.58
Clinical severity rating	.84
Specific phobia	
Situational fear	
Blood, injury, injection	.77
Situational	.73
Vomiting, choking, contracting an illness	.63
Animals	.64
Water	.54
Dental or medical procedures	.53
Situational avoidance	
Blood, injury, injection	.73
Situational	.73
Vomiting, choking, contracting an illness	.66
Animals	.72
Water	.48
Dental or medical procedures	.41
Clinical severity rating	.75
Major depression	
Key symptoms	.74
Clinical severity rating	.65
Dysthymia	
Key symptoms	.78
Clinical severity rating	.36
Any mood disorder (major depression or dysthymia)	
Clinical severity rating	.69

Note. ADIS-IV-L = Anxiety Disorders Interview Schedule for DSM-IV: Lifetime version. $N = 292$ for all analyses except for analyses of panic disorder/agoraphobia number of panic attacks, fear of panic attacks, and agoraphobic avoidance ratings ($ns = 97$). For all rs , $p < .001$.

and OCD without compulsions; Chorpita, Brown, & Barlow, 1998).

As in previous studies (Di Nardo et al., 1993; Williams et al., 1992), the current and lifetime diagnosis of DYS possessed poor reliability, further calling into question the utility of this category

as currently defined. Although the potential overlap of DYS and GAD is apparent (i.e., both disorders constitute chronic symptoms of negative affect), it is noteworthy that the vast majority of DYS disagreements involved other mood disorders. This was also true for MDD disagreements in which the anxiety disorders were rarely involved. This suggests that boundary issues within the mood disorders are a primary source of unreliability, often pertaining to limitations of the categorical approach such as differentiating (a) DYS from chronic MDD and (b) MDD and DYS from depressive disorder NOS. This also accounts for the findings of higher reliability when MDD and DYS were collapsed into one category than when they were analyzed as separate categories (see Tables 1 and 4).

Unreliability Due to Diagnostic Threshold Issues

Although a similar pattern of reliability estimates was obtained when any current diagnoses were examined (i.e., collapsing principal and additional diagnoses), interrater agreement of PD, OCD, and SPEC evidenced a marked decline relative to their estimates as principal diagnoses. Inspection of the sources of unreliability indicated that these categories were the most prone to disagreement involving diagnostic thresholds—that is, both interviewers recorded key features of the disorders but disagreed on the presence of sufficient impairment and distress to assign a formal DSM-IV diagnosis (e.g., this issue was responsible for 62% of SPEC disagreements). This was a strong contributing factor to reduced reliability of PD, OCD, and SPEC because additional diagnoses were more susceptible to the threshold issue than were principal diagnoses (i.e., by definition, a principal diagnosis is the disorder associated with the highest degree of distress or interference). Similarly, although excellent reliability was evident for the five SPEC types as principal diagnoses, these estimates declined for most SPEC types when collapsing principal and additional diagnoses. This again was attributable mainly to higher rates of diagnostic threshold disagreements, although certain SPEC types were more affected by this issue (i.e., animal, natural environment, blood/injury/injection); thus, defining the boundary of clinically significant interference and distress may be more difficult for some forms of SPEC (e.g., although marked impairment or distress may be clearly indicated in situational fears such as driving, it may be less apparent in fears of things such as animals, heights, etc., which the person rarely encounters or can avoid without considerable lifestyle impact).

The diagnostic threshold issue also illustrates the problem of measurement error introduced by imposing categorical cutoffs (i.e., DSM-IV criteria for the presence or absence of a disorder) on diagnostic features that operate largely in a continuous fashion (e.g., number, severity, and duration of symptoms and degree of distress). Evaluation of sources of unreliability suggests several other instances in which this occurred. Many of the diagnostic disagreements associated with GAD, MDD, and DYS involved anxiety disorder NOS and depressive disorder NOS. This indicates that both interviewers agreed on the presence of clinically significant features of the disorder in question (clinical severity ratings ≥ 4), but that one interviewer did not assign a formal anxiety or mood disorder diagnosis because of subthreshold patient report of the number or duration of symptoms (because of inconsistent report, change in clinical status, etc.). Another example of this

problem pertained to the severity specifiers for MDD. Whereas dimensional ratings of the severity of MDD features were quite reliable ($r = .74$; Table 5), the *DSM-IV* categorical specifiers of MDD severity evidenced poor reliability (κ s = .30 and .36; Table 2). Because of the measurement error, loss of information, and validity problems associated with the purely categorical approach to diagnostic classification in *DSM-IV*, researchers have called for incorporation of dimensional components in future nosological systems (e.g., Blashfield, 1990; Brown, in press; Frances, Widiger, & Fyer, 1990).

Indeed, favorable reliability was found for most composite dimensional ratings of disorder features and for single ratings such as the clinical severity ratings (Table 5). These findings are noteworthy in view of the wide use of these measures as indexes of treatment outcome (e.g., Borkovec & Costello, 1993; Brown & Barlow, 1995) and as indicators in studies of the nature of emotional disorders (e.g., Brown et al., 1998). Although intended to provide psychometric justification for composite scoring, the results of the factor analyses of the ADIS-IV-L dimensional ratings may have implications for the typology of some disorders. For example, analysis of fear ratings of 17 specific phobia situations did not support the presence of a distinct factor representing natural environment-type phobias. Instead, such fears either tended to be associated with situational fears (heights, storms) or failed to aggregate saliently with any other fear (water). This result could be interpreted to support prior arguments and preliminary findings that some natural-environment-type fears (e.g., heights) are better construed as situational-type phobias (Antony et al., 1997).

Summary and Conclusions

The current findings provide support for the reliability of most *DSM-IV* emotional disorders as assessed by the ADIS-IV-L and elucidate sources of error in the diagnosis of these conditions. However, these findings clearly show that the *DSM-IV* anxiety and mood disorders were differentially affected by the various sources of unreliability. Besides MDD and DYS (whose disagreements frequently involved each other), only GAD and SPEC had considerable rates of disagreements involving other diagnostic categories (mood disorders in GAD, agoraphobia in SPEC), which might suggest that these disorders are more prone to error associated with overlapping key or associated features. For many categories (e.g., SOC and OCD), disagreements rarely involved other disorders and were primarily due to problems in defining and applying a categorical threshold to the classification of the number, severity, or duration of symptoms (e.g., disagreements on clinical vs. subclinical diagnoses and disagreements involving NOS diagnoses). Although the clinical versus subclinical issue was less relevant in reliable diagnosis of PDA, GAD, MDD, and DYS, unreliability related to categorical threshold was evident in these disorders by the high incidence of disagreements with NOS diagnoses and MDD versus DYS. Thus, the high rate of disagreements involving thresholds and NOS diagnoses indicated that in many cases interviewers concurred on the presence of the features of a given disorder; however, unreliability was introduced through the difficulties in applying the *DSM-IV* categorical cutoff to these features. These data support the need for continued research that may ultimately and unequivocally document the importance of

dimensionally based assessment systems in improving our formal approaches to the classification of psychological disorders.

References

- Abel, J. L., & Borkovec, T. D. (1995). Generalizability of *DSM-III-R* generalized anxiety disorder to proposed *DSM-IV* criteria and cross-validation of proposed changes. *Journal of Anxiety Disorders, 9*, 303-315.
- 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.
- Antony, M. M., Brown, T. A., & Barlow, D. H. (1997). Heterogeneity among specific phobia types in *DSM-IV*. *Behaviour Research and Therapy, 35*, 1089-1100.
- Antony, M. M., Moras, K., Meadows, E. A., Di Nardo, P. A., Utech, J. E., & Barlow, D. H. (1994). The diagnostic significance of the functional impairment and subjective distress criterion: An illustration with the *DSM-III-R* anxiety disorders. *Journal of Psychopathology and Behavioral Assessment, 16*, 253-263.
- Baer, L. (1994). Factor analysis of symptoms subtypes of obsessive-compulsive disorder and their relation to personality and tic disorders. *Journal of Clinical Psychiatry, 55*, 18-23.
- Barlow, D. H., Brown, T. A., & Craske, M. G. (1994). Definitions of panic attacks and panic disorder in *DSM-IV*: Implications for research. *Journal of Abnormal Psychology, 103*, 553-564.
- Blashfield, R. K. (1990). Comorbidity and classification. In J. D. Maser & C. R. Cloninger (Eds.), *Comorbidity of mood and anxiety disorders* (pp. 61-82). Washington, DC: American Psychiatric Press.
- Borkovec, T. D. (1994). The nature, functions, and origins of worry. In G. Davey & F. Tallis (Eds.), *Worrying: Perspectives on theory, assessment, and treatment* (pp. 5-33). New York: Wiley.
- Borkovec, T. D., & Costello, E. (1993). Efficacy of applied relaxation and cognitive behavioral therapy in the treatment of generalized anxiety disorder. *Journal of Consulting and Clinical Psychology, 61*, 611-619.
- Brawman-Mintzer, O., Lydiard, R. B., Emmanuel, N., Payeur, R., Johnson, M., Roberts, J., Jarrell, M. P., & Ballenger, J. C. (1993). Psychiatric comorbidity in patients with generalized anxiety disorder. *American Journal of Psychiatry, 150*, 1216-1218.
- Brown, T. A. (1996). Validity of the *DSM-III-R* and *DSM-IV* classification systems for anxiety disorders. In R. M. Rapee (Ed.), *Current controversies in the anxiety disorders* (pp. 21-45). New York: Guilford Press.
- Brown, T. A. (in press). The classification of anxiety disorders: Current status and future directions. In D. J. Stein & E. Hollander (Eds.), *Textbook of anxiety disorders*. Washington, DC: American Psychiatric Press.
- Brown, T. A., & Barlow, D. H. (1992). Comorbidity among anxiety disorders: Implications for treatment and *DSM-IV*. *Journal of Consulting and Clinical Psychology, 60*, 835-844.
- Brown, T. A., & Barlow, D. H. (1995). Long-term outcome in cognitive-behavioral treatment of panic disorder: Clinical predictors and alternative strategies for assessment. *Journal of Consulting and Clinical Psychology, 63*, 754-765.
- Brown, T. A., Barlow, D. H., & Liebowitz, M. R. (1994). The empirical basis of generalized anxiety disorder. *American Journal of Psychiatry, 151*, 1272-1280.
- Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the *DSM-IV* anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology, 107*, 179-192.
- Brown, T. A., Marten, P. A., & Barlow, D. H. (1995). Discriminant validity of the symptoms constituting the *DSM-III-R* and *DSM-IV* associated

- symptom criterion of generalized anxiety disorder. *Journal of Anxiety Disorders*, 9, 317–328.
- Brown, T. A., Moras, K., Zinbarg, R. E., & Barlow, D. H. (1993). Diagnostic and symptom distinguishability of generalized anxiety disorder and obsessive-compulsive disorder. *Behavior Therapy*, 24, 227–240.
- Chorpita, B. F., Brown, T. A., & Barlow, D. H. (1998). Diagnostic reliability of the *DSM-III-R* anxiety disorders: Mediating effects of patient and diagnostician characteristics. *Behavior Modification*, 22, 307–320.
- Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology*, 100, 316–336.
- Costello, C. G. (1992). Research on symptoms versus research on syndromes: Arguments in favour of allocating more research time to the study of symptoms. *British Journal of Psychiatry*, 60, 304–308.
- Di Nardo, P. A., & Barlow, D. H. (1988). *Anxiety Disorders Interview Schedule—Revised (ADIS-R)*. Albany, NY: Graywind.
- Di Nardo, P. A., Brown, T. A., & Barlow, D. H. (1994). *Anxiety Disorders Interview Schedule for DSM-IV: Lifetime version (ADIS-IV-L)*. San Antonio, TX: Psychological Corporation.
- Di Nardo, P. A., Moras, K., Barlow, D. H., Rapee, R. M., & Brown, T. A. (1993). Reliability of *DSM-III-R* anxiety disorder categories using the Anxiety Disorders Interview Schedule—Revised (ADIS-R). *Archives of General Psychiatry*, 50, 251–256.
- Fleiss, J. L., Nee, J. C. M., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86, 974–977.
- Frances, A., Widiger, T., & Fyer, M. R. (1990). The influence of classification methods on comorbidity. In J. D. Maser & C. R. Cloninger (Eds.), *Comorbidity of mood and anxiety disorders* (pp. 41–59). Washington, DC: American Psychiatric Press.
- Mannuzza, S., Fyer, A. J., Martin, L. Y., Gallops, M. S., Endicott, J., Gorman, J. M., Liebowitz, M. R., & Klein, D. F. (1989). Reliability of anxiety assessment: I. Diagnostic agreement. *Archives of General Psychiatry*, 46, 1093–1101.
- Mannuzza, S., Schneier, F. R., Chapman, T. F., Liebowitz, M. R., Klein, D. F., & Fyer, A. J. (1995). Generalized social phobia: Reliability and validity. *Archives of General Psychiatry*, 52, 230–237.
- Marten, P. A., Brown, T. A., Barlow, D. H., Borkovec, T. D., Shear, M. K., & Lydiard, R. B. (1993). Evaluation of the ratings comprising the associated symptom criterion of *DSM-III-R* generalized anxiety disorder. *Journal of Nervous and Mental Disease*, 181, 676–682.
- Starcevic, V. (1995). Pathological worry in major depression: A preliminary report. *Behaviour Research and Therapy*, 33, 55–56.
- Stein, M. B., Walker, J. R., & Forde, D. R. (1994). Setting diagnostic thresholds for social phobia: Considerations from a community survey of social anxiety. *American Journal of Psychiatry*, 151, 408–412.
- Summerfeldt, L. J., Richter, M. A., Antony, M. M., & Swinson, R. P. (1999). Symptom structure in obsessive-compulsive disorder: A confirmatory factor-analytic study. *Behaviour Research and Therapy*, 37, 297–311.
- Turner, S. M., Beidel, D. C., & Stanley, M. A. (1992). Are obsessional thoughts and worry different cognitive phenomena? *Clinical Psychology Review*, 12, 257–270.
- Williams, J. B. W., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope, H. G., Rounsaville, B., & Wittchen, H. (1992). The Structured Clinical Interview for *DSM-III-R* (SCID): II. Multisite test-retest reliability. *Archives of General Psychiatry*, 49, 630–636.

Received February 7, 2000

Revision received August 1, 2000

Accepted August 3, 2000 ■