From diverse perspectives, there is little doubt that depressive symptoms cohere to form a valid and distinct syndrome. Research indicates that an evidence-based assessment of depression would include (a) measures with adequate psychometric properties; (b) adequate coverage of symptoms; (c) adequate coverage of depressed mood, anhedonia, and suicidality; (d) an approach to suicidality that distinguishes between resolved plans and preparations and desire and ideation; (e) assessment of the atypical, seasonal, and melancholic subtypes; (f) parameters of course and chronicity; and (g) comorbidity and bipolarity. These complexities need to be accounted for when certain assessment approaches are preferred, and when ambiguity exists regarding the categorical versus dimensional nature of depression, and whether and when clinician ratings outperform self-report. The authors conclude that no one extant procedure is ideal and suggest that the combination of certain interviews and self-report scales represents the state of the art for evidence-based assessment of depression.

Classic work on construct validity (Cronbach & Meehl, 1955) has shown that, before posing the question of “How do we measure something?,” we must first answer the question “What exactly is this something?” By implication, therefore, knowledge of what depression is must precede and guide endeavors to assess it. So what exactly is depression? The answer to that question is complex: There are some things we know about depression, and some things we do not know, and both sets of things inform evidence-based assessment of depression.

We summarize some daunting issues regarding the construct of depression and its assessment. Is depression a continuous or dichotomous phenomenon? Should self-report scales or structured clinical interviews be favored and, if so, under what conditions? How should one handle complex issues like course and chronicity, depression subtypes, and comorbidity? Insofar as possible, we attempt to answer these and other questions. We believe that evidence-based assessment of depression must grapple with complexities like (a) ambiguity regarding the categorical versus dimensional nature of depression; (b) ambiguity regarding whether and when clinician ratings outperform self-report; and (c) when certain approaches (e.g., self-report, interviews) are to be preferred over others. Moreover, our view is that evidence-based assessment of depression must include (a) adequate coverage of the symptoms that cohere to form the valid and distinct syndrome of major depression; (b) adequate coverage of depressed mood, anhedonia, and suicidality, which appear to be of particular importance; (c) adequate coverage of short-term change in symptoms (e.g., waxing and waning of symptoms as people undergo treatment); (d) regarding suicidality, an approach that distinguishes between resolved plans and preparations and suicidal desire and ideation; (e) assessment of the atypical subtype and, less definitively, the melancholic subtype as well; (f) parameters of course and chronicity; (g) comorbidity and bipolarity; and (h) assessment tools with adequate psychometric properties. We conclude that structured clinical interviews, supplemented by well-validated self-report scales represent the state of the art for evidence-based assessment of depression.

What We Do Not Know

Categorical or Dimensional?

Is depression a category, an all-or-none, “either-you-have-it-or-you-don’t” phenomenon? Alternatively, is depression a continuum, distributed along a graded dimension, similar to temperature or mass, where a thing can have very little of it or a little more of it, and so on, up to having very much of it? There are staunch advocates of both positions (e.g., Coyne, 1994; Vredenburg, Flett, & Krames, 1993).

This question has not been definitively answered and is, in essence, a taxometric question. The key idea of taxometrics is that...
the indicators of true taxa (i.e., all-or-none classes), as opposed to key indicators of continua, intercorrelate because they serve to distinguish taxa members from taxa nonmembers. Thus, in a mixed group of taxon members and nonmembers, the indicators will intercorrelate; however, in a pure subsample of taxon members or a pure subsample of taxon nonmembers, the indicators’ intercorrelation will be less (see Joiner & Schmidt, 2002; Waller & Meehl, 1998).

The value of taxometric analysis for addressing these types of nosologic questions has been highlighted in a series of studies on posttraumatic stress disorder (A. M. Ruscio, Ruscio, & Keane, 2002), worry (A. M. Ruscio, Borkovec, & Ruscio, 2001), psychopathy (Harris, Rice, & Quinsey, 1994), schizotypy (e.g., Lenzenweger & Korfine, 1992), dissociation (Waller, Putnam, & Carlson, 1996), comorbidity (Meehl, 2001), and eating disorders (Gleaves, Lowe, Snow, Green, & Murphy-Eberenz, 2000; see Watson, 2003, and Widiger, 2001, and for some criticisms of taxometrics, which, in the context of Meehl’s extensive body of work on the topic, are unpersuasive). Even so, the application of taxometrics to psychopathology is still in its relatively early stages. This is particularly true with regard to mood disorder nosology: Major depression has not been well studied using taxometric methods. Franklin, Strong, and Greene (2002), Haslam and Beck (1994), and Grove et al. (1987) are exceptions, but these studies did not assess major depression per se, they have some important methodological limitations (e.g., Franklin et al. used only three indicators, all self-report, and all from the Minnesota Multiphasic Personality Inventory [MMPI]), and they come to differing conclusions.

There are two additional taxometric studies on depression, and these stand out in terms of their quality (Beach & Amir, 2003; J. Ruscio & Ruscio, 2000). Ruscio and Ruscio reported analyses on two samples: In one sample, self-report items from depression scales were used as indicators; in the other, self-report MMPI items were used as indicators. Although there was some evidence of depression taxonicity in some of their data, they concluded that the majority of findings supported a dimensional view. This was an impressive study; however, it shares the limitation with other taxometric studies on depression (e.g., Franklin et al., 2002) of sole reliance on self-report items from depression scales. Beach and Amir reached different conclusions in their study. Crucially, they distinguished between symptoms of mere distress versus hallmark symptoms of major depression (cf. Santor & Coyne, 2001a) and found evidence for taxonicity of hallmark symptoms but not of general distress symptoms (Ambrosini, Bennett, Cleland, & Haslam, 2002, reported a similar finding in adolescents). This raises the possibility that depression is taxonic, but that most studies to date have not used sufficiently rigorous indicators to detect depression taxa.

What is needed is a study that matches the analytic sophistication of J. Ruscio and Ruscio (2000) and the general assessment approach of Beach and Amir (2003; see also Ambrosini et al., 2002; Haslam & Beck, 1994) but that also focuses on a large sample of people (“enriched” to ensure substantial subsets of individuals with mood disorders) and that uses multimodal assessment indicators (i.e., clinician ratings as well as depression self-report scales) focusing on hallmark symptoms of major depression. We eagerly await the results of such a study. Until then, the jury is out regarding whether depression is best viewed categorically or dimensionally. Therefore, evidence-based assessment of depression needs to be flexible enough to account for the fact that depression itself may be taxonic or continuous in nature and that different assessment goals may call for approaches that are taxonic (diagnosis as goal) or continuous (e.g., characterizing severity as goal).

Self-Report Versus Clinician Ratings?

Although a consensus appears to have emerged that clinician-rated, structured clinical interviews represent a kind of gold standard in psychopathology assessment (a view that we ourselves endorse, at least partially, later in this article), it is worth noting that the empirical basis for this consensus is not beyond question. A problem with this literature is the lack of studies in which both self- and clinician ratings of symptoms are referenced to a third data source that is relatively impervious to various biases associated with both self- and clinician ratings (e.g., neuroendocrine response to biological challenge). Moreover, there are instances in which patients’ self-ratings may outperform clinicians’ ratings, specifically in the area of suicidality. For example, Jobes, Jacoby, Cimbolic, and Hustead (1997) tracked patients who had initially presented with suicidal symptoms and, based on the later course of the suicidal symptoms, classified patients into acute resolver and chronic nonresolver outcome categories. Patients’ initial self-ratings were better than clinicians’ ratings at predicting eventual group membership. Similarly, among a large group of suicidal young adults, Joiner, Rudd, and Rajab (1999) reported that patients’ self-ratings were more telling than clinicians’ ratings regarding recurrence of suicidal symptoms at 6-, 12-, and 18-month follow-up sessions. The well-known biases and distortions associated with symptom self-report scales, taken together with instances in which self-report may have outperformed clinicians’ ratings, suggest two things: (a) Conclusions regarding this topic must be specific and nuanced (the view that one or the other approach is always categorically best does not withstand empirical scrutiny); and (b) in order to produce nuanced conclusions, more work with very powerful designs is needed.

Does Evidence-Based Assessment of Depression Possess General and Incremental Clinical Utility?

Meehl (1959) suggested that there are four levels of increasingly important incremental validity for assessment procedures. At the first level, assessment provides clinicians with reliable and valid information about the patient. At the second level, assessment provides clinicians with reliable and valid information that cannot be concurrently obtained from routine procedures in which the clinician would normally engage. At the third level, the test provides clinicians with reliable and valid information about patients earlier in time than would be available by some other means. Finally, at the fourth level, the test provides reliable and accurate information about patients, which then helps in treatment (for a general review of issues in treatment utility of assessment, see Hayes, Nelson, & Jarrett, 1987). We later discuss an evidence-based approach that reaches the first level, but to our knowledge there is not an adequate database to evaluate the other levels, although a case could be made that work on suicide assessment and atypical depression may reach some or all of the other levels. This area represents a key avenue for future research.
What We Know

Depressive symptoms—sadness; anhedonia; suicidality; slowing; low self-esteem—guilt; and problems with energy, concentration, sleep, and appetite—cohere to form a valid and distinct syndrome. The Diagnostic and Statistical Manual of Mental Disorders (4th ed.; DSM–IV; American Psychiatric Association, 1994) defines a major depressive episode as the experience of at least five of these nine symptoms (at least one of which must be sadness or anhedonia) most of the time for at least 2 weeks. The 2-week criterion represents a bare minimum; the average length of a major depressive episode is approximately 7 months (e.g., Shapiro & Keller, 1981).

The syndrome has very well-characterized parameters, including course, comorbidity patterns, and treatment response. From diverse perspectives, including epidemiological (Cross-National Collaborative Group, 1992; Kessler, Rubinow, Holmes, Abelson, & Zhao, 1997), genetic (Kendler et al., 1995), neurobiological (Post, 1992), therapeutic (Franchini, Gasperini, Perez, & Smeraldi, 1997), prognostic (Coryell et al., 1995), personological (Agrams, Rosendahl, Card, & Alexopoulos, 1994), psychological (Abramson, Metalsky, & Alloy, 1989), and others, the construct enjoys considerable and consistent support.

Depression Can Be Reliably and Validly Assessed

Attesting to its validity as a syndrome, a variety of depression assessment instruments with demonstrated psychometric properties (i.e., adequate reliability and validity) are available. With regard to reliability, appropriate assessment devices display internal consistency (the extent to which each item measures the same construct) and short-term test–retest stability (the extent to which a measure yields the same score across repeated administrations). Interview-based or clinician-rating measures further demonstrate interrater reliability (the extent to which different raters arrive at the same score for an individual). With regard to validity, appropriate assessment devices demonstrate construct validity, which includes content validity (the degree to which a measure represents a balanced and adequate coverage of depression), convergent validity (the extent that a measure correlates with other measures of depression), discriminant validity (the extent to which a measure does not assess constructs other than depression), and predictive validity (the extent to which a measure predicts people’s depression status at a later point in time).

We recommend the following psychometric criteria as a minimum standard for selecting depression assessment devices: (a) internal consistency and short-term test–retest reliabilities ≥ .70; (b) interrater reliability ≥ .70 for interview-based or clinician rating measures; (c) demonstrated construct validity; and (d) examination of the psychometric properties of the devices in at least two separate samples. All measures described here exceed these standards; we elaborate on their specific psychometric properties later.

Despite the extensive support for depression as a construct, including the existence of numerous valid measurement devices, there remain some complexities that need to be accounted for in evidence-based assessment of depression.

Depression Is More Than Just Distress or Negative Affect

To be sure, depressed people are distressed and experience negative affect, but distressed people are not necessarily depressed. In many studies using self-report depression inventories (which are suffused with negative affect content), this fact was neglected. Partly in reaction, two important lines of work evolved.

One line of work emphasizes the distinction between clinical forms of depression on the one hand, and distress or demoralization on the other hand. Santor and Coyne (2001a), for example, used a nonparametric item response model to address this point. They determined whether the probability of expressing an individual depressive symptom was different depending on whether a participant reached or did not reach threshold for a clinical diagnosis of depression. Participants who did and did not meet threshold were equated regarding overall severity of depressive symptoms. Despite this, between-groups differences on some symptoms emerged. Specifically, depressed mood, anhedonia, and suicidality were more likely to be expressed in individuals who met clinical threshold than in those who did not, whereas hypochondriasis and middle insomnia were more likely to be expressed in individuals who did not meet clinical threshold than in those who did. The authors interpreted these differences to be inconsistent with the view of depression as a simple continuum ranging from demoralization or distress to clear presence of severe disorder. We suggest that a main implication of the study by Santor and Coyne, as well as similar studies, is that key symptoms like anhedonia, sad mood, and suicidality need to be well assessed.

Anhedonia and sad mood also are highlighted in a second line of work on the suffusion of depression measures by negative affect. The tripartite model of depression and anxiety (Clark & Watson, 1991) argues that indicators associated with general negative affect (including sad mood) do not discriminate depression from anxiety very well, in that people with depression and anxiety (and other people still) all experience considerable amounts of general negative affect. By contrast, anhedonia is a good discriminator, in that those with depression often experience depression, whereas those with anxiety and other conditions do not (unless they are also comorbid for depression). The empirical work evaluating this model has been mostly supportive (e.g., Brown, Chorpita, & Barlow, 1998).

Yet the model has the potential to produce misunderstanding, specifically in underestimating sadness as an indicator of depression and overestimating anhedonia as a depression indicator. In a study of adolescents who were monitored into early adulthood, Lewinsohn, Pettit, Joiner, and Seeley (2003) found that depressed mood was the most commonly experienced symptom across all episodes of major depression. Indeed, 90% to 95% of all episodes were characterized by depressed mood. To a degree, this is only to be expected because the DSM requires that either depressed mood or anhedonia be present for a diagnosis of major depression. Even though anhedonia also is prioritized by the DSM, the percentages of major depressive episodes characterized by anhedonia in the Lewinsohn et al. study ranged from 65% to 76%, not as high as those associated with depressed mood.

Lewinsohn et al. (2003) also investigated the concordance of depressed mood and anhedonia. Of the 564 individuals with a history of major depression, 67% concurrently experienced both depressed mood and anhedonia in their first episode, 28% of
depressed individuals experienced depressed mood without anhedonia, and 5% experienced anhedonia without coexisting depressed mood. A similar pattern of results emerged for recurrences of depression beyond the first episode.

These findings suggest conditional probabilities that refine the tripartite model and that are relevant to empirically based assessment of depression. Specifically, given that one is experiencing a major depressive episode, the likelihood that depressed mood is a symptom is quite high and is higher than the likelihood that anhedonia is a symptom. Given that one is experiencing a major depressive episode characterized either by depressed mood alone or by anhedonia alone, the likelihood that the episode is characterized by depressed mood is approximately five times greater than the likelihood that it is characterized by anhedonia. These probabilities prioritize the importance of depressed mood and appear to contradict the tripartite model’s emphasis of anhedonia. However, the tripartite model is actually arguing for a different set of conditional probabilities. Specifically, the model asserts that, given that someone is experiencing depressed mood, the likelihood of a major depressive episode is not necessarily high because depressed mood is common in general and is a frequent associated feature of many mental disorders. By contrast, given that someone is experiencing anhedonia, the likelihood of a major depressive episode is elevated because anhedonia is not particularly common in general and is not as frequent an associated feature of many mental disorders. One exception is schizophrenia, but data suggest that those with major depression are characterized by anhedonia, even compared with those with schizophrenia (Joiner, Brown, & Metalsky, 2003).

To summarize, depression is clearly more than just distress, demoralization, or depressed mood, but Lewinsohn et al. (2003) show that depressed mood is the most common symptom of a major depressive episode, more so than anhedonia and other symptoms. The point of the tripartite model is that depressed mood, although very common among those experiencing major depression, is not very specific to the syndrome; anhedonia, by contrast, is more unique to major depression. Santor and Coyne’s (2001a) results are consistent with these conclusions and also emphasize a third symptom that is as important as depressed mood and anhedonia in evidence-based assessment of depression: suicidality.

Suicidality, a clear indicator of the depressive syndrome, is a complex construct in itself. Like others (e.g., Beck, Kovacs, & Weissman, 1979), we found that suicidality’s factor space can be adequately explained by two factors, which, although correlated (approximate r = .50), were distinguishable: resolved plans and preparations and suicidal desire and ideation (Joiner, Rudd, & Rajah, 1997).

The resolved plans and preparation factor included the following indicators: a sense of courage to make an attempt; a sense of competence to make an attempt; availability of means to and opportunity for attempt; specificity of plan for attempt; preparations for attempt; duration of suicidal ideation; and intensity of suicidal ideation. Those who endorse the items on this factor may experience more intense and acute forms of suicidality, and the foreboding tone of the items highlights the possibility that they may be at substantial risk for impending suicide attempt, a possibility supported by other studies (e.g., Joiner et al., 2003).

The suicidal desire and ideation factor was composed of the following symptoms: reasons for living (reverse scored), wish to die, frequency of ideation, wish not to live, passive attempt, desire for attempt, expectancy of attempt, lack of deterrents to attempt, and talk of death or suicide. The factor is made up of items that assess ongoing thoughts, ideas, and desires regarding suicide (perhaps chronic, less acute suicidal ideation) but not of items reflecting intense ideation or readiness to commit suicide.

The evidence-based assessment of suicidality is a topic in its own right (see, e.g., Cukrowicz, Wingate, Driscoll, & Joiner, 2004; Joiner, Walker, Rudd, & Jobes, 1999). A main implication of suicide assessment for depression assessment is that resolved plans and preparation is an extremely important indicator of dangerousness across categories of mental disorders, including but not limited to depression. By contrast, suicidal desire and ideation are more depressotypic. In Joiner et al. (1997), the partial correlation between suicidal desire and ideation and depressive symptoms as measured by the Beck Depression Inventory (BDI; Beck & Steer, 1987), controlling for the resolved plans and preparation factor, was .36 (p < .01), whereas the partial correlation between resolved plans and preparation and BDI depressive symptoms, controlling for the suicidal desire and ideation factor, was .06 (ns). The presence of suicidal symptoms corresponding to either factor is of clinical concern (which should be handled according to systematic procedures; e.g., Joiner, Walker, et al., 1999), but the symptoms of suicidal desire and ideation are a common indicator of depression, whereas resolved plans and preparation is relatively less specific to depression and of more concern regarding dangerousness.

**There Are At Least Two Valid Depression Subtypes and Maybe More**

In 1586, Timothy Bright wrote that “when any conceit troubleth you that hath no sufficient grounde of reason, but riseth only upon the frame of your brayne, which is subject unto the humor, that is right melancholike” (Bright, 1586/1969, p. 194). Bright was arguing for an endogenous subtype of depression, unrelated to external stressors. Although interest in subtyping has waxed and waned, often in step with nosologic and therapeutic innovations (e.g., interest and controversy peaked following Kraepelin’s seminal work and then ebbed somewhat until another peak after the introduction of antidepressants and electroconvulsive therapy [ECT]), some evidence has accrued for the existence of an endogenous–melancholic subtype, although the originally key idea of lack of external stressors or causes has not been well supported. DSM’s melancholic subtype includes symptoms such as loss of interest in all or almost all activities, lack of reactivity to pleasurable stimuli, distinct quality of depressed mood, symptoms regularly worse in the morning, early morning awakening, weight loss, excessive guilt, and psychomotor retardation (American Psychiatric Association, 1994). It is important to note that the lack of reactivity symptom is different than the notion that symptoms were caused exclusively endogenously. The symptom refers to lack of reactivity to pleasure once depressed, not to lack of external causes. In fact, recent work indicates that almost all depressive episodes are judged by careful interviewers to be reactive in nature (i.e., with presence of a precipitating event; e.g., Lewinsohn et al., 2003).

Studies in support of the existence and validity of this subtype include the factor analytic (i.e., a melancholic–endogenous symptom factor is demonstrated to exist; e.g., Andreasen, Grove,
Maurer, 1980; Mendels & Cochrane, 1968), the taxometric (i.e., melancholic–endogenous depression is a naturally occurring taxon; Haslam & Beck, 1994), the neurochemical (i.e., decreased brain dopamine consumption in melancholic–endogenous depression; van Praag & Korf, 1971), the neurohormonal (e.g., patients with melancholic–endogenous features break away from dexamethasone suppression; Zimmerman, Coryell, Pfohl, & Stangl, 1985), the therapeutic (e.g., melancholic–endogenous depression predicts good response to ECT; Abou-Saleh & Coppen, 1983), and the prognostic (i.e., melancholic–endogenous features predict good treatment outcome; Copeland, 1983). This pattern of findings led Rush and Weissnerburger (1994), in their summary of melancholic–endogenous features and DSM–IV (American Psychiatric Association, 1994), to conclude that the concept has considerable validity and clinical utility.

However, with regard to each type of study mentioned earlier, there are sound studies that do not support the validity of the endogenous subtype. For example, with regard to the dexamethasone suppression test, K. B. Miller and Nelson (1987) found that it was not a reliable discriminator between depressed patients with and those without melancholic–endogenous features. Berger (1982) obtained a similar result and also reported that other neurophysiological markers (e.g., sleep architecture, urinary free cortisol) did not discriminate patients with endogenous versus non-endogenous depressions. Family studies do not support the hypothesis that endogenous depression is more familial than non-endogenous depression (Andreasen et al., 1986).

Support for the atypical subtype of major depression appears to be somewhat more uniformly positive (see Quitkin, McGrath, Stewart, & Klein, 2003, for a discussion of some key issues). DSM’s atypical subtype includes symptoms such as the presence of mood reactivity to positive events, significant weight gain, hypersomnia, leaden paralysis (arms or legs feeling like they are weighted down), and interpersonal rejection sensitivity (APA, 1994). Several studies have shown atypical depression to be discernible from other depressive symptoms factorially (e.g., Kendler et al., 1996) and neurophysiologically (see Klein, 1993). Perhaps the most persuasive line of evidence comes from studies on differential treatment response: People with atypical depression have been shown to be particularly responsive to monoamine oxidase inhibitors such as phenelzine (e.g., Klein, 1993).

Seasonal affective disorder (SAD) is another depression subtype for which validity data are compelling (e.g., Lam & Levitan, 2000). SAD is characterized by occurrence at times of the year when sunlight is less abundant (i.e., in the winter). The disorder is thus latitude dependent (i.e., occurs more frequently at more northern latitudes). Its symptoms include usual major depression symptoms (e.g., sadness–irritability; inability to enjoy things; sleep, appetite, and concentration disturbance; suicidal thoughts), except that oversleeping and overeating are more common than insomnia and appetite loss, respectively. Therefore, any assessment device that is equipped to measure usual depression symptoms (including oversleeping and overeating) is applicable to SAD. Specific devices are available to assess seasonality, the best studied of which is the Seasonal Pattern Assessment Questionnaire (Rosenthal, Genhardt, Sack, Skwerer, & Wehr, 1987; see also Young, Blodgett, & Reardon, 2003). This instrument, together with usual depression assessment measures, allows evidence-based assessment of this depression subtype.

Course and Chronicity as Defining Parameters of Depression

Several aspects of depression course and chronicity can be distinguished, each of which is a negative prognostic indicator and thus deserves attention in assessment: early age of first onset (e.g., before age 13); severity of past episode; past episode characterized by psychotic or severe suicidal symptoms; an episode superimposed on a preexisting dysthymia (i.e., “double depression”; Keller & Shapiro, 1982); long duration of an individual episode; resumption of symptoms shortly after the remission of an episode (i.e., relapse); reestablishment of major depression after a diagnosis-free period (i.e., recurrence; Mueller et al., 1999); and long-standing, residual, subclinical depressive symptoms that persist after an episode (Judd & Akiskal, 2000). The waxing and waning of symptoms as people undergo treatment also represent an important course-related consideration, and self-report severity scales appear appropriate in this context.

Depressions that first occur in later life, compared with those that first occur in childhood, adolescence, and early adulthood, may be less severe (Burvill, Hall, Stumpfer, & Emmerson, 1989), less associated with suicidal symptoms (Cassano, Akiskal, Savino, & Soriani, 1993), and less related to personality problems (Abrams, et al., 1994). They also may be about equally common in men and women (whereas earlier onset depressions are more common in women; Krishnan, Hays, Tulder, & George, 1995), less associated with first-degree relatives’ depression risk (Bland, Newman, & Orn, 1986), and more related to neurologic or medical disease (Alexopoulos, Young, Meyers, & Abrams, 1988).

Comorbidity

Major depression has high rates of comorbidity with anxiety, eating, substance use, and personality disorders. Perhaps the highest rates of comorbidity are with the various anxiety disorders. The reasons for this overlap are interesting to consider because they likely differ per the different anxiety disorders. The reason for the overlap of major depression and generalized anxiety disorder (GAD) is likely shared cause; the genetic risks for major depression and for GAD overlap considerably or perhaps even entirely (Kendler, Neale, Kessler, & Heath, 1992). Accordingly, those with major depression should be screened for GAD and vice versa, because individuals who meet criteria for one are at high genetic risk for the other. The reasons for the association of major depression and obsessive–compulsive disorder (OCD), for example, are likely different. Shared etiology is unlikely (e.g., the basal ganglia are implicated in OCD but not in major depression; e.g., Rapoport, 1990); rather, those who have OCD suffer distress and life impairment to the degree that a major depressive episode is likely. In other words, depression is a common consequence of OCD (but

---

1 Abramson et al. (1989) proposed the theory-based subtype of hopelessness depression, with symptoms that overlap considerably with DSM major depression. Generally, this subtype is theoretically the product of the expectation of negative outcomes and the belief that one is powerless to affect such outcomes, which in turn are the products of stable and global attributions of the causes of negative events. The subtype has some validity data behind it (e.g., Joiner et al., 2001), but to our knowledge there are no data relevant to its clinical utility.
OCD is not a common consequence of depression). Therefore, those with OCD should be screened for depression, but those with depression are not particularly likely to have OCD.

**Bipolarity**

The usually obvious difference between bipolar spectrum disorders and major depression should be emphasized. It is not unusual for depressed people who do not have bipolar disorder (but, e.g., have heard of it through friends or the media) to describe their non-ill times as manic episodes. However, careful questioning, perhaps in the context of structured clinical interviews, reveals that these episodes are periods of normal functioning (good, stable mood; ideas more or less in tune with reality; sleeping between 6 and 9 hr per night). These times may seem manic to a formerly depressed person because they contrast so starkly with the symptoms of major depression. A useful rule of thumb, applicable in the majority of cases (but not all), is that people with bipolar spectrum disorders will be able to identify three phases: a manic phase, a depressed phase, and a euthymic, essentially symptom-free phase. (One exception to this rule involves people who are so ill that they are virtually always depressed or manic, but their severity of illness makes differential diagnosis from major depression a moot point.) By contrast, people with major depression can identify two major phases: depressed versus not. (Again, this is not a perfect rule because, e.g., people sometimes experience long-standing, residual, subclinical depressive symptoms that persist after a major depressive episode.)

Another point of differentiation involves sleep. People with major depression very much desire to sleep well and often point to insomnia as among the most troubling of their symptoms. By contrast, people with bipolar disorder experience reduced need and desire for sleep and are often engaged in goal-directed activity when they would otherwise be asleep.

**Ethnicity, Gender, and Age**

Overall, there is no clear evidence that depression assessment measures contain inherent biases regarding ethnicity and gender (see Blazer, Landerman, Hays, Simonsick, & Saunders, 1998; Gallo, Cooper-Patrick, & Lesikar, 1998; Iwata, Turner, & Lloyd, 2002, for representative work). With regard to age, there is some concern that older people may obtain inflated scores on self-report depression measures as a result of endorsement of somatic and vegetative symptoms that stem from nondepressive sources (e.g., medical problems). It is worth noting, therefore, that assessment measures specifically developed for geriatric populations exist (e.g., the Geriatric Depression Scale; Yesavage et al., 1983). However, there is also evidence that scales such as the BDI and the Center for Epidemiologic Studies Depression Scale (CES–D) also possess adequate properties in geriatric samples (Lewinsohn, Seeley, Roberts, & Allen, 1997; Olin, Schneider, Eaton, & Zemansky, 1992).

**Purpose of Assessment Drives the Selection of Measures**

As mentioned, numerous valid assessment devices exist for depression. Just as current knowledge of depression informs approaches to evidence-based assessment, the purpose of the assessment guides the selection of appropriate assessment tools. Diagnosis is the primary goal of many assessments, and we have devoted considerable attention to that topic. Other objectives of assessment include the preliminary identification of those who may meet diagnostic criteria for a depressive disorder (i.e., screening), treatment planning, and the monitoring of symptom change during treatment. We focus here on two of these purposes of assessment: screening and symptom monitoring.

Screening, or initial assessment to determine the presence of depressive symptoms, is used primarily to identify individuals at high and low risk of a depressive disorder (e.g., Alloy et al., 2000). Unlike structured clinical interviews, which are time intensive and costly, measures appropriate for screening are characterized by rapid administration and scoring. They provide a quick estimate of the number and severity of depressive symptoms, and their primary advantage is in the minimal outlay of resources. The validity of certain instruments for screening purposes has been supported, including self-report measures such as the CES–D (Radloff, 1977; Roberts, Lewinsohn, & Seeley, 1991) and BDI–II (Beck, Steer, & Brown, 1996; Kumar, Steer, Teitelman, & Villacis, 2002). These screening instruments are limited, however, in that they do not assess the degree of functional impairment required to meet diagnostic criteria for depressive disorders and are restricted with regard to the course of symptoms. As such, screenings that indicate the presence of depressive symptoms should be followed by further evaluation before diagnosis and intervention.

Another purpose of depression assessment is monitoring of symptom severity, including change in symptoms over time. Three issues must be addressed when considering measures for symptom monitoring: pragmatic concerns, sensitivity to change, and treatment context. Some measures and strategies are more pragmatic than others for symptom monitoring. We address this in greater detail later. However, we note here that structured clinical interviews are appropriate, valid, and reliable diagnostic instruments but are not practical for assessing symptom severity and change over multiple occasions. Although they provide extensive assessment data, repeated administration would prove to be an unnecessarily arduous task (especially in some settings, e.g., medical). Rather, self-report measures such as the BDI–II or brief clinician rating scales such as the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1967; I. W. Miller, Bishop, Norman, & Maddever, 1985) are preferable.

In addition to pragmatic constraints, a measure’s sensitivity to change is of high importance in monitoring depressive symptoms. Evidence suggests that clinician rating measures are more sensitive to symptom changes than self-report measures, although the accuracy of the heightened sensitivity has been questioned (Lambert & Lamberts, 1999).

The third issue pertaining to symptom monitoring is treatment setting. As symptom severity and course vary across inpatient, outpatient, and nonclinical settings, so should symptom-monitoring strategies. For example, Clark, Cook, and Snow (1998) considered several depression symptom assessment strategies across depressed psychiatric inpatients, chronic medically ill patients, and normal controls. They found that negative cognitive symptoms of depression distinguished psychiatric inpatients from chronically ill medical patients. Clark et al. observed that clinician rating measures such as the HRSD were also better able to distinguish somatic and behavioral discriminators of depression. Thus,
in the context of inpatient psychiatric settings and general medical settings, clinician rating measures may be most appropriate for tracking changes in depressive symptoms. However, given that certain items on such measures are only endorsed at extremely severe levels of depression (e.g., weight loss, depersonalization; Santor & Coyle, 2001b), their appropriateness for monitoring symptoms may be limited to treatment contexts of severe depression (i.e., inpatient settings). Self-report measures, such as the BDI–II, may be preferable for tracking symptom change in less severe cases (i.e., outpatient settings).

Based on the Foregoing, Which Measures?

Based on the foregoing, an evidence-based assessment of depression would include the following: (a) selection of measures with adequate psychometric properties; (b) adequate coverage of the symptoms that cohere to form the valid and distinct syndrome of major depression; (c) adequate coverage of depressed mood, anhedonia, and suicidality, which appear to be of particular importance; (d) regarding suicidality, an approach that distinguishes between resolved plans and preparations and suicidal desire and ideation; (e) assessment of the atypical and seasonal subtypes and, less definitively, the melancholic subtype; (f) parameters of course and chronicity; and (g) comorbidity and bipolarity. Also, several complexities need to be considered and accounted for: (a) when certain assessment approaches are to be preferred over others, such as when the purpose is screening, diagnosis, or monitoring of symptom change, respectively; (b) ambiguity regarding the categorical versus dimensional nature of depression; and (c) ambiguity regarding whether and when clinician ratings outperform self-report.

We suggest that certain structured clinical interviews meet all or nearly all of these criteria. The Structured Clinical Interview for the DSM–IV (SCID; First, Spitzer, Williams, & Gibbon, 1995) is perhaps the most commonly used structured clinical interview. It is a comprehensive interview used to make Axis I diagnoses for the DSM–IV (e.g., major depressive episode), and as such, it closely conforms to the DSM–IV diagnostic decision trees. Regarding the criteria outlined above, the general reliability and validity of the SCID are well established (e.g., see Rogers, 2001, for more specific reliability and validity information on this and other roughly similar structured clinical interviews like the Diagnostic Interview Schedule [Robins, Helzer, Croughan, & Ratcliff, 1981] and the Schedule for Affective Disorders and Schizophrenia [Endicott & Spitzer, 1978]). In addition, its coverage of depressed mood, anhedonia, suicidality, and other symptoms is good, and interrater reliability per each of the syndrome’s symptoms is adequate (e.g., Lewinsohn et al., 2003). Regarding suicidality, the SCID differentiates between a specific plan or attempt on the one hand (cf. resolved plans and preparations) and thoughts of death and suicidal ideation on the other (cf. suicidal desire and ideation), albeit in abbreviated form. The SCID specifically assesses for the atypical subtype. Although many of the symptoms of the melancholic subtype are evaluated as well, some are not; seasonality is not well assessed. Parameters of course and chronicity, such as age of onset, number of past episodes, and preexisting dysthymia, are noted. Comorbidity and bipolarity are very well assessed. The SCID was specifically designed to formally assign depressive (and other diagnoses), but it is too time and effort intensive to serve well as a mass screening tool. It is also not ideal to track changes in symptoms over time. Regarding the categorical versus dimensional nature of depression, the SCID is designed to produce categorical diagnoses, but counts of specific symptoms can add rough dimensional information. Regarding clinician ratings versus self-report, the SCID’s format requires some degree of clinical judgment but also includes a number of open-ended queries; the combination of open-ended queries and clinical judgment may incorporate the advantages of both self-report and clinician ratings.

The MINI International Neuropsychiatric Interview (MINI; Sheehan et al., 1998) is a highly structured and relatively brief clinical interview designed to assess current DSM–IV diagnoses of all major Axis I disorders, including major depression, bipolar disorders I and II, as well as other disorders. Regarding the criteria outlined above, the MINI does not have as extensive a track record of general reliability and validity as the SCID but shows good agreement with other standardized structured clinical interviews for Axis I disorders (Sheehan et al., 1997). The MINI’s coverage of depressed mood, anhedonia, suicidality, and other symptoms is good. Regarding suicidality, the MINI differentiates between a specific plan or attempt on the one hand (cf. resolved plans and preparations) and thoughts of death and suicidal ideation on the other (cf. suicidal desire and ideation), albeit in abbreviated form. The MINI specifically assesses for the melancholic subtype but not the atypical subtype. Parameters of course and chronicity are not well evaluated, although whether or not past episodes occurred is noted. Comorbidity and bipolarity are well assessed by the MINI. In addition, the MINI was specifically designed to formally assign depressive and other diagnoses, but it is too time and effort intensive to serve well as a mass screening tool. It is also not ideal for tracking changes in symptoms over time. Regarding the categorical versus dimensional nature of depression, the MINI is designed to produce categorical diagnoses, but counts of specific symptoms can add rough dimensional information. Regarding clinician ratings versus self-report, the MINI’s format requires clinical judgment and does not include many open-ended queries; it may, therefore, incorporate the advantages of clinician ratings but not self-report.

Overall, the SCID has the advantage over the MINI of detailed coverage of disorders, a more extensive track record, coverage of the atypical subtype, more comprehensive coverage of course and chronicity parameters, and incorporation of the advantages of clinician rating and self-report. The MINI, alternatively, is briefer and easier to learn and use and provides good coverage of the melancholic subtype. Combination of the MINI’s melancholic subtype module with the SCID would meet all the criteria outlined previously, except for assessment of seasonality, use in mass screening, and the tracking of changes in symptom severity over time.

The Longitudinal Interval Follow-Up Evaluation (LIFE; Keller et al., 1987) also deserves mention. It is similar in format to the SCID but is geared toward detailed information about the course of psychiatric symptoms and disorders since an initial diagnosis, with rigorous criteria for recovery from a disorder (i.e., symptom free for 8 or more weeks). It is thus particularly strong on the parameters of course and chronicity, especially as they are affected by treatment. However, for the short-term tracking of symptom changes, including in response to treatment, as well as mass
screening, self-report scales, which are mentioned next, may be preferable.

In contrast to the structured clinical interviews, self-report symptom scales struggle to meet some of the criteria outlined previously. On the other hand, they may be particularly apt for tracking short-term (e.g., weekly) vacillations in symptom severity (their usual designed purpose) as well as mass screening. For both of these purposes, a time- and effort-intensive assessment is not workable. Self-report scales, which are brief and can be simultaneously administered to large groups, serve these purposes well.

The BDI–II is a 21-item self-report inventory of depressive symptoms. Each item is rated on a scale ranging from 0 to 3; total scores thus range from 0 to 63. The BDI–II is a revision of earlier versions of the BDI (Beck & Steer, 1987) that conforms to changes made in the DSM–IV (American Psychiatric Association, 1994). Regarding the criteria outlined previously, the internal consistency and factor structure of the BDI–II has received ample support among outpatient samples of adults and adolescents (coefficient alphas typically at or above .90; e.g., Beck et al., 1996; Steer, Ball, Ranieri, & Beck, 1997, 1999; Steer, Kumar, Ranieri, & Beck, 1998; Steer, Rissmiller, & Beck, 2000), indicating that the BDI–II is very reliable and well validated as an index of depressive symptom severity. The BDI–II includes items to assess sadness, anhedonia, and suicidal ideation. Regarding suicidality, the BDI–II fails to assess symptoms of resolved plans and preparations. In contrast to the earlier BDI, the BDI–II assesses certain symptoms of the atypical subtype of major depression, such as hypersomnia and weight gain, but is not appropriate alone for diagnosing subtypes. Parameters of course and chronicity as well as comorbidity and bipolarity are not evaluated. The BDI–II is appropriate for use in mass screening situations to identify those who might need intensive assessment. Because the BDI–II is brief and its metric is well known, we view it as the premier instrument for the assessment of depressive symptom severity and the tracking of short-term changes in severity in outpatient settings (e.g., in response to treatment). It was designed for this purpose and was specifically designed not to establish formal depressive diagnoses. Regarding the categorical versus dimensional nature of depression, the BDI–II is designed primarily as a dimensional measure of symptom severity, and cut scores for categorical placement are relatively arbitrary. Regarding clinician ratings versus self-report, its format incorporates the advantages of self-report ratings but does not allow for clinical judgment.

The CES–D is a self-report index of the frequency of occurrence of 20 depressive symptoms, each rated on a scale ranging from 0 to 3, yielding total scores with a possible range from 0 to 60. Regarding the criteria outlined previously, the CES–D has been demonstrated to be a reliable and valid assessment instrument in a variety of samples (Joiner, Pfaff, & Acres, 2002; Roberts et al., 1991). It assesses depressed mood and anhedonia but fails to evaluate suicidality. The CES–D does not assess subtypes of major depression. Parameters of course and chronicity as well as comorbidity and bipolarity are not evaluated. Because it was designed for use with general community samples, the CES–D is perhaps the best instrument for mass screening situations. It has received less empirical attention than the BDI–II as an instrument to monitor symptom changes and was not designed to establish formal depressive diagnoses. Regarding the categorical versus dimensional nature of depression, the CES–D is designed primarily as a dimensional measure of symptom severity, and cut scores for categorical placement are relatively arbitrary. Regarding clinician ratings versus self-report, the CES–D’s format incorporates the advantages of self-report ratings but does not allow for clinical judgment.

The Inventory to Diagnose Depression (IDD; Zimmerman & Coryell, 1987) performs similarly to the BDI–II and CES–D with regard to most criteria. Reliability and validity data for the scales are not as extensive as for the BDI and CES–D but appear to be adequate (Uehara, Sato, Sakado, & Kameda, 1997; Zimmerman & Coryell, 1987). The primary distinguishing feature of the IDD is its performance with regard to parameter and course. In particular, the IDD includes questions on chronicity and history. One version of the scale is geared toward current symptoms, and for each item respondents are asked to rate not only its severity but also its duration (i.e., whether it has lasted more or less than 2 weeks). A second version of the scale assesses the most depressed week in one’s life. Here again, respondents are asked not only to rate its severity but also whether it has lasted more or less than 2 weeks. Moreover, respondents are asked about total length of any depressive episode, impairment resulting from the episode, and total number of past episodes. Both versions include 22 items, each rated on a scale ranging from 0 to 4 and each yielding total score ranges from 0 to 88. Given the emphasis on parameters, like chronicity and impairment, the IDD is clearly the instrument of choice for attempts to formally diagnose people using self-report. Nevertheless, even studies supporting the IDD in this regard sound a note of caution about diagnosing using self-report (e.g., Uehara et al., 1997), a caution with which we agree.2

A final assessment device that merits discussion is the HRSD. This is neither a structured interview nor a self-report inventory; rather, it was developed primarily as a clinician rating scale of depression severity for patients already diagnosed with a depressive disorder (Rehm & O’Hara, 1985). HRSD versions of different item length are available (e.g., 17, 21, 24), although only the 17-item version is typically used for assessing depression severity. Regarding the criteria outlined above, the interrater reliability, internal consistency, and construct validity of the HRSD have received ample support (e.g., Bech, 1987; Rehm & O’Hara, 1985). The HRSD includes items to assess sadness, anhedonia, and suicidality. Regarding suicidality, the HRSD fails to assess symptoms of resolved plans and preparations. The HRSD includes a subscale for assessing the melancholic subtype but does not assess the atypical or seasonal subtypes. Parameters of course and chronicity and comorbidity and bipolarity are not evaluated. The HRSD may be used for individual or small group screenings but is not well suited for use in mass screening situations. Similar to the BDI–II, the HRSD is most appropriate for assessing depressive symptom

---

2 Other self-report scales are available as well. For example, Gross, Keyes, and Greene (2000) concluded that Scale 2 of the MMPI and MMPI–2 had moderate accuracy in assessing depression but noted that the Depression content scale did not add much beyond Scale 2. Wetzel, Khadivi, and Moser (1998) found that single MMPI scales were not very accurate in classifying depression but regression-based composites of scales were better. MMPI–2 composites were better than MMPI composites in this study. Because the MMPI–2 is lengthy, it is not optimal for one of self-report’s main uses in depression assessment: repeated checking of symptom severity. However, if the goal is to screen people for depression as well as many other potential conditions, the MMPI–2 has utility.
severity and tracking short-term changes in severity, particularly in inpatient settings. Regarding the categorical versus dimensional nature of depression, the HRSD is designed primarily as a dimensional measure of symptom severity, and cut scores for categorical placement are relatively arbitrary. Regarding clinician ratings versus self-report, the HRSD relies heavily on clinical judgment and, therefore, may incorporate the advantages of clinician ratings but not self-report.

Conclusion

On the basis of available knowledge, our recommendation for the optimal evidence-based depression assessment in clinical settings includes (a) the SCID to establish formal mood disorder diagnoses; (b) the MINI module on melancholic features to supplement the SCID; (c) the Seasonal Pattern Assessment Questionnaire (Rosenthal et al., 1987) to assess for the possibility of a seasonal component to any diagnosed mood disorder; (d) the BDI–II to assess severity of depressive symptoms and short-term change in depressive symptoms; and (e) the LIFE to formally assess remission of disorder once BDI–II scores are stably low for many weeks. For mass screening situations, the CES–D and BDI–II seem to be wise choices.

From diverse perspectives, there is little doubt that depressive symptoms cohere to form a valid and distinct syndrome, with several well-characterized parameters. Consideration of these parameters, such as the prominence of depressed mood, anhedonia, and suicidality as symptoms, factors of suicidality, subtyping, course and chronicity, and so forth, shows that thorough evidence-based assessment of depression requires intensive and somewhat complex procedures. No one extant procedure is ideal, but the combination of measures described previously appears to cover all key issues with minimal disadvantages and represents the state of the art for evidence-based assessment of depression.

References


