Cognitive Therapy and Research, Vol. 11, No. 3, 1987, pp. 289-299

Issues and Recommendations Regarding Use of the Beck Depression Inventory

Philip C. Kendall[1]
Temple University

Steven D. Hollon
Vanderbilt University

Aaron T. Beck
University of Pennsylvania

Constance L. Hammen
University of California, Los Angeles

Rick E. Ingram
San Diego State University

Issues concerning use of the Beck Depression Inventory (BDI) for the self-report of depressive symptomatology are raised and considered. Discussion includes the stability of depression and the need for multiple assessment periods, specificity and the need for multiple assessment measures, and selection cut scores and the need for terminological accuracy. Recommendations for the continued use of the BDI, designed to facilitate the integration of diverse studies and improve research on self-reported depression, are provided.

KEY WORDS: depression; assessment; Beck Depression Inventory (BDI); research recommendations.

[1] Address all correspondence to Philip C. Kendall, Department of Psychology, Temple University, Philadelphia, Pennsylvania 19122.

Depression is a widely used term, and, in part as a result of its colloquial presence, there is an associated ambiguity. At one end of the spectrum, depression refers to a clinical syndrome that can include psychotic features, somatic disturbance, psychomotor retardation, and risk for suicide. At the other end of the spectrum, depression is a commonly used term to denote "feeling a little down." For clinical research, the meaning of depression is more than simple semantics—different definitions and uses of the label "depressed" have important implications.

The professional use of the term *depression* has several levels of reference: symptom, syndrome, nosologic disorder (Beck, 1967; Lehmann, 1959). Depression can itself be a symptom—for example, being sad. As a syndrome, depression is a constellation of signs and symptoms that cluster together (e.g., sadness, negative self-concept, sleep and appetite disturbances). The syndrome of depression is itself a psychological dysfunction but can also be present, in secondary ways, in other diagnosed disorders. Finally, for depression to be a nosologic category careful diagnostic procedures are required during which other potential diagnostic categories are excluded. The presumption, of course, is that a discrete nosologic entity will ultimately prove to be etiologically distinct from other discrete entities, with associated differences likely in course, prognosis, and treatment response. Although two individuals might evidence the same manifest symptomatology with respect to the syndrome of depression, learning that one also showed all the hallmarks of paranoid schizophrenia, while the other did not, would lead us to exclude the first individual from the nosologic category of primary major depressive disorders. The Beck Depression Inventory (BDI) is a sensitive measure of syndrome depression, but it was never intended to be a nosologic screening device.

A number of studies investigating a wide range of potentially psychopathological processes rely on self-report psychometric devices such as the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961; Beck, 1967) to define experimental groups of depressed individuals. The relevance of results obtained with these subjects to differentially defined depressed (nosologic category) samples is unclear. While we believe that these diverse studies do provide important findings to the literature and can help contribute to a meaningful understanding of the phenomenon, there are potential shortcomings as well. The present paper is intended to discuss relevant issues and provide reasonable guidelines to help overcome certain limitations and to facilitate clear and unambiguous communication regarding research on depression.

STABILITY OF DEPRESSION AND THE NEED FOR MULTIPLE ASSESSMENT PERIODS

Recent studies have shown clearly that despite relatively high coefficients of test-retest reliability, over 50% of those who initially score above some defined cutoff criterion change classification when retesting is conducted within hours (Hatzenbuehler, Parpal, & Matthews, 1983), days (Zimmerman, 1986), or 1 to 4 weeks (Deardorff & Funabiki, 1985; Hammen, 1980). Many change from some classification of depressed to nondepressed, while others change from more to less depressed. The overall stability coefficients appear to be contributed largely by relatively nondepressed scorers. One implication of the issue of instability, raised by Sacco (1981), is that the experimenter bears the burden of proof in demonstrating that the status of participants at selection and experimental testing is the same. We concur with Sacco's (1981) recommendation that the presence of depressive affect must be verified at the time of any experiment. In most cases, subjects should be assessed during at least two time periods: once at initial selection and once immediately preceding the experiment (if they are indeed different). This methodology actually addresses two points: stability over time and emotional status at the time of experimental procedures. That is, multiple assessment not only increases confidence in the depressed condition of the subjects but also ensures that they are experiencing such affect at the time of the actual research. Research based solely on a single-administration BDI classification cannot be generalized to the full clinical state, since most such nosological classifications require some minimal duration of distress.[2]

SPECIFICITY AND THE NEED FOR MULTIPLE ASSESSMENT METHODS

The specificity issue, as related to the BDI, is largely an issue of construct validity (Cronbach & Meehl, 1955). A construct is a concept that is

[2] There are three options here: (1) a one-shot assessment of BDI and dependent variables, (2) assessing BDI at one time and dependent variables at another, and (3) assessing BDI at one point and reassessing BDI along with dependent variables at a second point. It is the second design that is especially troublesome since any dependent variable collection occasion should be accompanied by a reassessment of the subject's depression levels. If one wants to generalize to nosologic depression, then the third design (with structured interviews) seems preferred. If one is solely interested in the presence of syndromal depression, then the third design is best, but the first design would be satisfactory.

"constructed," on the basis of scientific information, to explain and organize aspects of existing knowledge. A psychological instrument is said to have construct validity only when a number of studies are integrated and judged to support the underlying theoretical construct.

The issue of psychopathological specificity concerns both (a) high-end specificity; the degree to which high scores on the BDI are associated solely with increased levels of depression and not other disorders and (b) low-end specificity; the extent to which the BDI assesses a continium ranging from no depression to severe depression, with low scores indicating the absence of psychopathology.

High-End Specificity

A concern for the use of the BDI is the specificity of the scale for the assessment of "depression" as a homogeneous psychopathological entity. Recent data are consistent in suggesting that the instrument is relatively sensitive but only moderately specific to a single nosologic category. Several investigators have shown that in community samples (e.g., Oliver & Simmons, 1984) as well as among college students (Deardorff & Funabiki, 1985; Hammen, 1980), significant proportions of high scorers do not have diagnosable disorders, and of those who do have disorders, many do not have current primary affective disorders. The BDI can be viewed favorably as a measure of syndromal depression, but BDI scores alone are insufficient as indices of nosologic depression. Recalling that the BDI was designed to be a measure of *syndrome,* not *nosologic,* depression, a major portion of this nonspecificity is hardly surprising, since the syndrome is supposed to be present in nosologic primary affective disorders but need not be absent in other nosologic entities (e.g., the schizophrenias, anxiety disorders, substance abuse disorders). However, elevated syndrome depression scores can also reflect a variety of dispositional qualities, responses to stressful life events or other transitory happenings that reflect no known nosologic entity. Research has found high correlations between BDI scores and other self-reported symptoms (e.g., Gotlib, 1984). Watson and Clark (1984) argued for a construct of "negative affectivity" as a term for the general tendency to experience and report discomfort, and display negative views of the self and situations, encompassing what has been termed trait anxiety, ego strength, neuroticism, and the like.

Evidence attests to the fact that depressive and anxious affect are correlated in any given sample (e.g., Hollon & Kendall, 1980, reported a correlation .79 between depression and generalized anxiety in a sample of college students). This association has important interpretive implications; results obtained from a "depressed" sample may be a reflection of depressive affect, other negative affect, or some interaction among the different affective states. Indeed, recent research by Ingram, Kendall, Smith, Donnell, and Ronan (in press) found that depressed college students evidenced one pattern of automatic thoughts while test-anxious subjects evidenced a different pattern. However, subjects who were *both* depressed and test-anxious evidenced still a different pattern. It would be potentially misleading to report that a certain pattern of psychopathological features is characteristic of depression when it may be characteristic of other disorders as well. This type of problem is particularly likely when depression is assessed solely at the syndromal level, since many different nosologic categories may evidence high levels of syndrome depression.

The specificity sword cuts both ways. While depression researchers must be aware of the potential confounding effects due to the presence of other negative affects, so must researchers who study anxiety or other affective states. Researchers who have relied on psychometric subject selection of a specific psychopathological state have typically not acknowledged this potential affective overlap and its implications. Future work would benefit greatly from the inclusion of assessments of other affective states as well as depression and analysis of the results to determine what effects are specific to depression relative to other mood conditions. The burden of proof must fall on the investigator to demonstrate the generalizability of the sample to particular clinical groups, or show that the results may be applied specifically to depression rather than to nonspecific distress conditions or other nosological categories.

Use of the BDI in conjunction with other assessments, the hallmark of the classic multimethod measurement methodology, promises to provide the most convincing diagnostic statements. Such an approach would be required in efforts to assess depression as a nosologic category. Use of multiple methods of assessment is desirable, but use of multiple assessments that are all syndromal in nature would not resolve this dilemma. For instance, use of the BDI and a Hamilton Depression Rating Scale (Hamilton, 1960), a clinician-rated instrument, might increase one's confidence in the accuracy of the syndromal assessment but would still not be sufficient for assigning a nosologic diagnosis. The simple inclusion of other self-report syndromal measures (e.g., the MMPI-D or the Zung) would be even less helpful in this regard, since they don't even vary the method of assessment, much less the level.

Multiple method measurements may be undertaken in conjunction with the need for multiple assessment periods and employ a multiple gate strategy. The multiple gate approach, as applied to the present topic (taken from epidemiological research) involves the administration of a self-report inventory, (e.g., BDI) and a selection of potential cases for study using a set-point score. A second assessment period is arranged for those subjects who qualify, and they are again administered scales along with a clinical, preferrably structured, diagnostic interview. The second self-report score must also meet a criterion set point and the subject must also qualify as depressed on the basis of the ratings emerging from the structured interview. Although the phrase *multiple gate* may be new, many psychotherapy outcome researchers follow such procedures. It is consistent with earlier discussions (e.g., Meehl & Rosen, 1955) of the use of a less accurate but less expensive screening test to identify people for some later, more time-consuming, but ultimately more accurate diagnostic test.

The multiple gate strategy has its greatest relevance when selecting cases from clinical settings where the base rate for other disorders is higher than, say, from an undergraduate sample. The structured diagnostic interview may be less essential when conducting an analogue study since there is a lesser likelihood of the existence of other diagnostic categories. For example, in a clinical setting in which 30% of the patients carry a diagnosis of schizophrenia, and half those patients also exhibit a clinical level of syndrome depression, the odds of selecting someone with a diagnosis of schizophrenia if one screens only with the BDI are much greater than in a analogue population in which the base rate for schizophrenia is only 1%. Thus, given that nosologic depression is a rather high base-rate disorder (with an estimated prevalence of about 10%), simple syndromal screening may be a relatively safe procedure in an analogue sample. The same procedure is, however, far more likely to produce a sample that is nosologically heterogeneous in a clinical setting and should clearly be supplemented by diagnostic interviewing in that context.

Low-End Specificity

Studies often use control groups composed of nondepressed individuals to compare to their depressed groups. While this is an appropriate strategy, the selection of such groups on the basis of low BDI scores can at times be problematic. For instance, procedures in which the nondepressed group is selected on the basis of having minimal scores on the BDI (O's and l's) may lead to inferential errors since this cell would probably be composed of "Pollyannas," professional daredevils, incipient hypomanics, and the kind of people who want to talk to you when you sit next to them on an airplane. Indeed, Hammen (1983) has noted that some individuals selected for very low levels of depressive affect (e.g., a score of 0 on the BDI) may be characterized by other forms of psychopathology (e.g., psychopathy, hypomania) rather than, or in addition to, the absence of depression. Use of these subjects as a "normal" comparison group is thus inappropriate. We recommend that researchers select control groups for being statistically "average" in level of depression, or that three groups be selected; one high in syndrome depression, one at the mean for syndrome depression, and one low in syndrome depression.

CUTTING SCORES AND THE NEED FOR GREATER TERMINOLOGICAL PRECISION

Variations in the cutting scores (for initially selecting depressed cases and those used to reflect clinical improvement) used in research with the BDI create the potentially troublesome situation in which different studies can not be readily compared and/or integrated. It may be desirable to establish a minimum set point in order to have greater comparability between studies, but as a practical matter, it is difficult to suggest universal set points since the problems of instability and heterogeneity occur when both 10 and 16 are used to identify relatively depressed samples (e.g., Deardorff & Funabiki, 1985; Hammen, 1980). The adoption of a higher cut score would surely reduce false positives at the expense of false negatives (e.g., Oliver & Simmons, 1984) but would not ensure stability or eliminate heterogeneity. Most researchers would probably care more about reducing false positives and might find that adoption of higher set points increases the likelihood of finding significant predicted effects. Nevertheless, caution must be used in interpreting and generalizing from such outcomes.

While we are not inclined to offer a rigid recommended cut score for different degrees of syndromal depression or for clinically meaningful change, we do suggest consideration of the following stance. Large sample psychometrics for the BDI typically evidence a skewed distribution with a mean in the area of 4 to 6. The range of scores from 0 to 9 may be viewed as normal, but it must be noted that selecting subjects whose scores are at the extremely low end may produce a comparison group with other real abnormalities unrepresentative of true normalcy. Mild levels of depression are associated with BDI scores of 10 to 20, with 10 to 17 suggesting dysphoria and greater than 17 more closely associated with depressive states. Scores of 20 to 30 reflect moderate depression, and scores greater than 30 reflect severe depression. Studies that have selected outpatients on the basis of meeting Feighner, RDC, or DSM-III criteria for primary unipolar depression typically report mean BDI scores around 30 (e.g., Beck, Hollon, Young, Bedrosian, & Budenz, 1985; Murphy, Simons, Wetzel, & Lustman, 1984; Rush, Beck, Kovacs, & Hollon, 1977). Consistent with these parameters, clients with BDI scores of 0 to 9 after clinical intervention may be reasonably viewed as remitted (e.g., Rush et al., 1977)[3]

An issue related to set points along the continuum of BDI scores concerns the continuity hypothesis. Are changes in measured level of depression matched by similar changes in some potentially covariant construct of interest? Or, stated differently, as one moves up the units of the BDI (e.g., specific scores or set point cutoffs), is there necessarily corresponding movement along another dimension of interest? For example, we know that psychomotor retardation is found in at least some severe depressives. Would we necessarily expect a little bit of psychomotor retardation in the presence of a little bit of "depression"? Continuity becomes particularly critical when we use analogue studies to test causal models; if modest levels of syndrome depression are relatively detectable (easily measured) and modest levels of some purported causal factors are less sensitively detected, then we could erroneously reject the hypotheses of covariation.

To the extent that depression is a continuous phenomenon, then a study that contrasts a group very low in measured depression with one only moderately low in depression might indeed yield valid inferences about the phenomenon. This would be equivalent to contrasting retardates and dull normals on performance in vocational school and finding that greater intelligence facilitates learning. While one might not be comfortable calling either group intelligent, most would be fairly comfortable talking about the role of intelligence on the basis of such a design. Similarly, we can readily imagine an analogue study contrasting normal nondepressives and subjects with moderate elevations on syndrome depression. While we might not want to label either group "depressed," we would still be justified in talking about the relationship of depression to other variables of interest, to the extent that the continuity hypothesis held true (see, however, Ruehlman, West, & Pasahow, 1985, for a provocative discussion of the complexities introduced when the continuity hypothesis does not appear to hold).

Base rates will influence the acquired mean scores on the BDI and consequent diagnostic decisions. Patients who score at the high end of the BDI and who came from the settings where persons with other disorders congregate (inpatient or outpatient clinical settings, all-night diners, self-help groups) would not be properly selected for depression or categorized as depressed using simply cut scores on the BDI. The odds for having elevated scores on the BDI for reasons stemming from causal processes different from those

[3] Use of short forms of the BDI is not recommended. Short forms may evidence significant correlations with the long form and nonsignificant differences between short and long form means, but the percentage of accurate classification of cases using a short form is typically much lower than with the long form (see also discussion in Kendall, in press).

Table I

Syndromal assessment (BDI Score)	Label
0–9	Nondepressed
10–15	Dysphoric
16–above, and Nosologic assessment	
Meets no criteria	Dysphoric
Meets criteria for affective disorders and nothing else	Depressed
Meets criteria for affective disorders and other predominant disorder (e.g., schizophrenia)	Other disorder with secondary depression

working in depression would simply be too great. Samples of college students will produce dramatically lower mean scores than samples of outpatient cases. When certain cut scores are implemented with college students, (e.g., scores in the 10–15 range) it seems wise to refer to them as dysphoric, unless detailed assessments using other measurement methods are employed (e.g., structured clinical interviews). Only when there is consistent agreement in the assessment of depression across measurement methods (e.g., BDI scores beyond a cut score, passing criteria on a structured clinical interview) would one be safe in assuming nosological classification and secure in using the label depressed (see Table I). Thus, unless it is demonstrated through the use of clinical interviews that BDI-evaluated samples meet diagnostic criteria, it might be desirable to avoid the term *depression* in favor of a term like *dysphoric* that implies nonspecific negative affectivity.

SUMMARY RECOMMENDATIONS

On the basis of the above considerations, then, we would put forward the following recommendations:

1. Assessments of depression should be made at a time concurrent with other assessments to which covariation is to be examined. Any initial screening for depression should be repeated on the day of subsequent testing. Multiple screenings reduce false positives that may occur owing to transitory distress.
2. Multiple method assessment is strongly recommended before employing the label "depressed" and implying a nosologic category. Multiple method assessments where the different methods tap both

syndromal and nosologic depression are preferred for studies purporting to investigate diagnosed depression.

3. The term *depression* should probably be reserved for individuals with BDI scores over 20 and preferably with concurrent diagnoses established by structured clinical interviews. Subjects selected solely on the basis of BDI scores should probably be referred to as "dysphoric."

We do not consider these recommendations to be either exhaustive or binding. Rather, we see them as helpful suggestions intended to facilitate the clarity of communication across the field and to protect the quality of the inferences that can be drawn from empirical data.

REFERENCES

Beck, A. T. (1967). *Depression.* Philadelphia: University of Pennsylvania Press.

Beck, A. T., Hollon, S. D., Young, J. E., Bedrosian, R. C., & Budenz, D. (1985). Treatment of depression with cognitive therapy and anitriptyline. *Archives of General Psychiatry, 42,* 142-148.

Beck, A. T., Steer, R. A., & Garbin, M. G. (in press). Psychometric properties of the BDI: 25 years later. *Clinical Psychology Review.*

Beck, A. T., Ward, C. M., Mendelson, M., Mock, J. E., & Erbaugh, J. K. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561-571.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Deardorff, W. W., & Funabiki, D. (1985). A diagnostic caution in screening for depressed college students. *Cognitive Therapy and Research, 9,* 277-284.

Gotlib, I. H. (1984). Depression and general psychopathology in university students. *Journal of Abnormal Psychology, 93,* 19-30.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23,* 56-62.

Hammen, C. L. (1980). Depression in college students: Beyond the Beck Depression Inventory. *Journal of Consulting and Clinical Psychology, 48,* 126-128.

Hammen, C. L. (1983). *Cognitive and social processes in bipolar affective disorders: A neglected topic.* Paper presented at the convention of the American Psychological Association, Anaheim, California.

Hatzenbuehler, L. C., Parpal, M., & Matthews, L. (1983). Classifying college students as depressed or nondepressed using the Beck Depression Inventory: An empirical analysis. *Journal of Consulting and Clinical Psychology, 51,* 360-366.

Hollon, S. D., & Kendall, P. C. (1980). Cognitive self-statements in depression: Development on an Automatic Thoughts Questionnaire. *Cognitive Therapy and Research, 4,* 383-395.

Ingram, R. E., Kendall, P. C., Smith, T. W., Donnell, C., & Ronan, K. (in press). Cognitive specificity in emotional distress. *Journal of Personality and Social Psychology.*

Kendall, P. C. (in press). Behavioral assessment and methodology. In G. T. Wilson, C. M. Franks, P. C. Kendall, & J. Foreyt. *Behavior Therapy in Review.* New York: Guilford.

Lehmann, H. J. (1959). Psychiatric concepts of depression: Nomenclature and classification. *Canadian Psychiatric Association Journal Supplement, 4,* S1-S12.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 191-216.

Murphy, G. E., Simons, A. D., Wetzel, R. D., & Lustman (1984). Cognitive therapy and pharmacotherapy: Singly and together in the treatment of depression. *Archives of General Psychiatry, 41,* 33-41.

Oliver, J. M., & Simmons, M. E. (1984). Depression as measured by the DSM-III and the Beck Depression Inventory in an unselected adult population. *Journal of Consulting and Clinical Psychology, 52,* 892-898.

Ruehlman, L. S., West, S. G., & Pasahow, R. J. (1985). Depression and evaluative schemata. *Journal of Personality, 53,* 46-92.

Rush, A. J., Beck, A. J., Kovacs, M., & Hollon, S. D. (1977). Comparative efficacy of cognitive therapy and pharmacotherapy in the treatment of depressed outpatients. *Cognitive Therapy and Research, 1,* 17-37.

Sacco, W. P. (1981). Invalid use of the Beck Depression Inventory to identify depressed college-student subjects: A methodological comment. *Cognitive Therapy and Research, 5,* 143-147.

Watson, D., & Clark, L. A. (1984). Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin, 96,* 465-490.

Zimmerman, M. (1986). The stability of the revised Beck Depression Inventory in college students: Relationship with life events. *Cognitive Therapy and Research, 10,* 37-44.