Psychological Assessment
2004, Vol. 16, No. 2, 120–132

Copyright 2004 by the American Psychological Association
1040-3590/04/$12.00 DOI: 10.1037/1040-3590.16.2.120

Reliability and Validity of the Beck Depression Inventory—II With Adolescent Psychiatric Inpatients

Augustine Osman, Beverly A. Kopper, and Frank Barrios
University of Northern Iowa

Peter M. Gutierrez
Northern Illinois University

Courtney L. Bagge
University of Missouri—Columbia

This investigation was conducted to validate the Beck Depression Inventory—II (BDI–II; A. T. Beck, R. A. Steer, & G. K. Brown, 1996) in samples of adolescent psychiatric inpatients. The sample in each substudy was primarily Caucasian. In Study 1, expert raters ($N = 7$) and adolescent psychiatric inpatients ($N = 13$) evaluated the BDI–II items to assess content validity. In Study 2, confirmatory factor analyses of several first-order solutions failed to provide adequate fit estimates to data for 205 boys, 203 girls, and the combined sample. Exploratory factor analyses identified new item-factor solutions. Reliability estimates were good (range = .72 to .91) for the BDI–II total and scale scores. In Study 3 ($N = 161$ boys and 158 girls from Study 2), preliminary evidence for estimates of concurrent, convergent, and discriminant validity were established for the BDI–II.

The Beck Depression Inventory—II (BDI–II; Beck, Steer, & Brown, 1996) is intended to address concerns with previous versions of the scale and to update the diagnostic criteria on which it is based (i.e., *Diagnostic and Statistical Manual of Mental Disorders,* 4th ed.; *DSM–IV;* American Psychiatric Association, 1994). Like previous editions, the BDI–II measures symptoms of depression severity yielding a total score by summing ratings for all 21 items (range = 0 to 63). The 4-point rating scale format (0 to 3) was retained for each item, with only a few modifications made in the wording of the original response options. Several (except 3 items) of the amended Beck Depression Inventory (BDI–IA) *item contents* (see Beck & Steer, 1993) were rewritten, and 4 old items (Items 14, 15, 19, and 20) were replaced by four new items (Items 11, 14, 15, and 19) "that represent the *DSM–IV* depression criteria" (Beck et al., 1996, p. 4). The time frame for rating each item was also modified from 1 week to 2 weeks to correspond to the *DSM–IV* diagnostic conceptualization for the major depressive disorders.

Although this instrument was intended for use with both adolescents and adults, the BDI–II manual does not report information regarding the reliability estimates, factorial validity, and convergent validity of the BDI–II for adolescent psychiatric inpatients. Furthermore, it is not clear in the manual whether the content validation processes of the BDI–II items included ratings by assessment experts on dimensions such as *specificity* (i.e., the extent to which the BDI–II items are unique to the assessment of depression severity) and *relevancy* (i.e., the extent to which the BDI–II items are identified as symptoms of a major depressive episode). To address these problems, the present studies were designed to explore several psychometric properties of the BDI–II, including content validity, factor structure, reliability estimates, and concurrent validity in adolescent inpatient samples. We followed several recommendations in establishing these psychometric characteristics (see Floyd & Widaman, 1995; Haynes, Richard, & Kubany, 1995).

Evidence of factorial validity for the BDI–II based on 500 adult psychiatric outpatients has been reported (Beck et al., 1996). However, using the same factor analytic and rotation procedures, Beck and colleagues failed to replicate the original item-factor compositions of the BDI–II in a small nonclinical sample of undergraduates. Results of studies in the manual assume that the two-factor oblique solution of the BDI–II (Somatic–Affective and Cognitive) is valid among adolescent psychiatric inpatients, even though these assumptions have received limited empirical validation. In addition, the majority of factor analytic studies have included adults only (e.g., Arnau, Meagher, Norris, & Bramson, 2001; Osman et al., 1997; Steer, Ball, Ranieri, & Beck, 1999). Only two studies have explored the factor structure of the BDI–II that included data from adolescent samples. Results of both studies did not successfully replicate the items that make up the separate

Augustine Osman, Beverly A. Kopper, and Frank Barrios, Department of Psychology, University of Northern Iowa; Peter M. Gutierrez, Department of Psychology, Northern Illinois University; Courtney L. Bagge, Department of Psychology, University of Missouri—Columbia.

We express our sincere thanks to all the expert raters, including those who did not wish to be listed in the acknowledgments. We also thank the inpatient raters, the study participants, and the staff at the Mental Health Institute for their support of this project. We thank Brian P. O'Connor for providing us with the software program for conducting the parallel analyses (using both the principal-components analysis and principal-axis factor methods).

Correspondence concerning this article should be addressed to Augustine Osman, Department of Psychology, University of Northern Iowa, 334 Baker Hall, Cedar Falls, IA 50614-0505. E-mail: augustine.osman@uni.edu

depressive disorder dimensions reported for the test-development samples.

Specifically, Dozois, Dobson, and Ahnberg (1998) evaluated the factor structure of the BDI–II and its comparability with the original BDI (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), using a subset of data ($n = 511$) from an independent sample of 1,022 Canadian undergraduate psychology students, ages 17 to 50 years.[1] These researchers failed to replicate clearly the item-factor compositions reported for the test-development samples in the manual. The authors subsequently used techniques of confirmatory factor analysis (CFA) with data from their second subsample ($n = 511$) of college students to compare their two-factor solution with the two-factor structure reported in the manual. These researchers concluded that the two-factor structure reported in the BDI–II manual provided only moderate fit estimates to their sample data. This study is relevant to the present investigations because the authors included data from 17- to 18-year-old adolescents. We note, however, that they did not evaluate separately the responses of adolescents and adults.

Steer, Kumar, Ranieri, and Beck (1998) identified three correlated factor solutions of the BDI–II in a sample of adolescent psychiatric outpatients, ages 12 to 18 years (105 boys and 105 girls). An iterated principal-axis factor analysis with promax rotation of the sample data revealed three factors, based on the scree plot criterion. The first factor (Cognitive) was composed of eight items (Items 2, 3, 7, 8, 9, 13, 14, and 19), the second factor (Somatic–Affective) also was composed of eight items (Items 1, 4, 12, 15, 16, 17, 18, and 20), and the third factor (Guilt/Punishment) was composed of three items (Items 5, 6, and 10). Item 10 also loaded highly (.45) on Factor 2. Items 11 and 21 did not load adequately (i.e., a preestablished criterion, by the authors, of .35 or greater) for retention on any of the extracted factors. The authors also extracted a second-order structure because of the moderate to high (i.e., values $\geq$ .30) intercorrelations observed among the first-order factors (range = .43 to .67). CFA procedures have not been used to examine the validity of this three-factor oblique solution among adolescent inpatients. Thus, the present study addresses this issue.

Regarding internal consistency of the BDI–II, the manual reported high alpha estimates (i.e., values $\geq$ .90) for this instrument in the adult clinical and nonclinical samples. Dozois et al. (1998) reported similar high alpha estimates for the BDI–II in the combined nonclinical sample of men ($\alpha = .92$) and women ($\alpha = .91$). To date, only two studies have been conducted with adolescent psychiatric inpatients, and in both studies alpha estimates of .90 or higher were reported for the BDI–II scores (Krefetz, Steer, Gulab, & Beck, 2002; Kumar, Steer, Teitelman, & Villacis, 2002). The present study also contributes to the BDI–II literature by providing mean interitem correlations (values of .15 or higher as considered good) as additional useful indices of internal consistency in an adolescent psychiatric sample (see Clark & Watson, 1995).

Evidence for the convergent and discriminative validity of the BDI–II was also examined in the studies with adolescents noted above. Specifically, in Krefetz et al.'s (2002) study ($N = 100$; ages = 12–17 years), the correlation between the BDI–II and the Reynolds Adolescent Depression Scale (Reynolds, 1987) was high ($r = .84$), suggesting good evidence of convergent validity for the mean BDI–II total score. Regarding discriminative validity, scores on both the BDI–II and the Reynolds Adolescent Depression Scale discriminated between adolescents who did and did not meet the *DSM–IV* criteria for a major depressive disorder.

Kumar et al. (2002) conducted a similar investigation of the discriminant validity of the BDI–II in an adolescent psychiatric inpatient sample ($N = 100$; age = 12–17 years). These researchers reported high and good discriminant validity indices for the BDI–II total (area under the curve [AUC] = .92) scores and their rationally defined Cognitive (AUC = .90) and Noncognitive (AUC = .90) subscales. A total cutoff score of 21 or higher was reported by the authors to have the highest clinically relevant positive predictive value (.85) and negative predictive value (.83). However, these researchers did not report the pattern of responses on the BDI–II for different diagnostic groups of adolescent inpatients (e.g., adolescents with a diagnosis of conduct disorder), as suggested in the BDI–II manual.

Specifically, the BDI–II manual contains descriptive data for individuals diagnosed with a range of adult psychiatric disorders and suggests that descriptive data (means and standard deviations) be reported for patients ages 13 and older as well. Thus, another purpose of this investigation was to report descriptive data on the BDI–II for adolescents with diverse *DSM–IV* diagnoses. Furthermore, because research suggests that gender differences in the expression of depressive disorder symptoms during adolescent years tend to be notable beginning at age 13 years (e.g., Twenge & Nolen-Hoeksema, 2002), we evaluated separately the structure of the BDI–II for boys and girls.

Taken together, the major objectives of this investigation were to provide much needed psychometric information about a scale designed to be used with adolescents and to replicate the only existing evidence that the BDI–II has good convergent and discriminant validities in adolescent psychiatric samples (Krefetz et al., 2002; Kumar et al., 2002). Information from existing studies is inadequate to support the use of the BDI–II with adolescent psychiatric inpatients, despite assurances in the manual that it is appropriate for individuals as young as 13.

Study 1: Content Validity Analyses

The major goal of Study 1 was to evaluate the content validity of the BDI–II. The manual noted that "the BDI–II was developed especially to assess the depressive symptoms listed as criteria for depressive disorders in the *DSM–IV*" (Beck et al., 1996, p. 25). Thus, in evaluating issues of content validity, we keyed several of the ratings (e.g., relevancy) to the *DSM–IV* major depressive disorder symptoms as a useful evaluative standard. The BDI–II is copyrighted; thus, the general aim of this study was to identify potential items that could be considered for modifications in future investigations.

[1] Because the BDI–IA items were modified in the development of the BDI–II, future studies might attempt to replicate Dozois et al.'s (1998) methods using the BDI–II and the BDI–IA items. The present study focused on findings related to the BDI–II in Dozois et al., not the original BDI.

Method

Participants

Expert rater sample. The expert raters (doctoral-level clinical psychologists)[2] were recruited because of their range of experiences and interests in assessment with children and adolescents. Of the 10 expert raters who were invited to participate, 7 returned the completed questionnaire packets. Each expert rater reported at least 5 years of experience in clinical assessments or research.

Adolescent psychiatric inpatient sample. The adolescent inpatients included 7 boys and 6 girls, ages 13–17 years (M = 15.00, SD = 1.53). The participants in this and subsequent studies (i.e., Studies 2 and 3) were recruited from the child and adolescent long-term care units of a state psychiatric hospital. The children's unit admits children ages 7 to 13 years, and the two adolescent units admit youths ages 14 to 17 years. In addition to the traditional therapeutic services, the hospital provides regular educational and psychoeducational programs for all the children and adolescents. All participants were enrolled in regular educational classes at the time of data collection for each study.

The Full Scale Wechsler Intelligence Scale for Children—Third Edition (Wechsler, 1991) IQs for this sample ranged from 90 to 112 (M = 100.46, SD = 6.81), indicating average-level intellectual functioning. Each potential participant was identified by the multidisciplinary treatment team members, at the request of Augustine Osman. Reviews of the medical chart diagnoses and also the social and developmental histories by Augustine Osman showed that each participant was assigned a primary *DSM–IV*-diagnosis of major depressive disorder by age 11 years.

Procedure

The expert raters were mailed a questionnaire packet that included an informed-consent form, the BDI–II, a brief demographic information questionnaire, a stamped and self-addressed return envelope, and the following: (a) a list of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., rev.; *DSM–IV–TR;* American Psychiatric Association, 2000*)* criteria for a major depressive episode, (b) a 5-point rating scale (1 = *not at all relevant* to 5 = *extremely relevant*) for rating the relevancy of each BDI–II item as a major depressive disorder symptom (*relevancy ratings*), and (c) a 5-point rating scale (1 = *not at all specific* to 5 = *extremely specific*) for rating each item as a symptom that is specific to major depressive disorder (*specificity ratings*). An overall BDI–II rating scale was also included for the raters to provide additional comments and other relevant ratings (see the Appendix).

The inpatient adolescents provided informed consent or assent and completed a brief demographic questionnaire and the following: (a) a 5-point rating scale (1 = *very hard to read and understand* to 5 = *extremely easy to read and understand*) to rate the extent to which each item (i.e., group of the BDI–II statements) was easily read and understood (*clarity ratings*) and (b) a 5-point rating scale (1 = *not at all useful* to 5 = *extremely useful*) to rate the statement "How well will each group of statements correspond to what you would say (i.e., the statement is *useful* to you) when you talk to a mental health professional about how you feel?"

The research protocol for conducting this study was approved by both the university Institutional Review Board and the hospital's program review committee. For the adolescent inpatients, parental or legal guardian consents were obtained during intake sessions as approved by the research review committees.

Results

Expert Rater Sample

Table 1 shows the results of the relevancy and specificity ratings of the BDI–II items. The overall mean relevancy ratings ranged from 2.90 to 4.81 (M = 4.14, SD = 0.64). The relevancy ratings for the individual BDI–II items ranged from 2.71 (Item 21) to 4.71 (Items 4 and 14). For the specificity ratings, the overall mean ratings ranged from 2.52 to 4.24 (M = 3.51, SD = 0.72). The specificity ratings for the individual BDI–II items ranged from 2.50 (Item 21) to 4.50 (Item 9).

When asked to list specific BDI–II items that are considered inappropriate for use with adolescents, 4 experts listed Item 21 (loss of interest in sex); 2 experts listed Item 13 (indecisiveness) and Item 19 (concentration difficulty); 2 experts listed Item 6 (punishment feelings); 1 expert listed Item 3 (past failure), Item 11 (agitation), and Item 17 (irritability); and 1 expert listed Item 7 (self-dislike) and Item 8 (self-criticalness) as inappropriate items.

Additionally, 5 experts indicated "yes," 1 indicated "no," and 2 indicated "not sure" to the question "Does the BDI–II contain all the essential symptoms seen in adolescents who are clinically diagnosed with major depressive disorder?" A majority of the experts indicated (M = 4.0) that the BDI–II total score was "very useful" to the assessment of the severity of major depressive disorder symptoms for adolescents. Regarding the overall adequacy ratings of factor structure (see the Appendix, Item 1), 4 experts indicated "yes," 1 indicated "no," and 2 indicated "not sure" to the question regarding the adequacy of the two-factor structural dimension of the BDI–II.

Furthermore, additional comments made by the experts regarding the overall adequacy of the BDI–II included the need to (a) include more behavioral items, (b) include other assessment instruments whenever the BDI–II is used with adolescents, (c) use developmentally appropriate language for some of the items (symptoms), (d) include items that may be differentiating in terms of depression versus adolescent developmental "turmoil," and (e) recognize that the BDI–II items can be "faked (good/bad) quite easily." One expert rater further noted that the *DSM–IV–TR* criteria for major depressive episode should be expanded to include six or more symptoms over 3 or 4 weeks.

Adolescent Psychiatric Inpatient Sample

The overall mean clarity ratings ranged from 3.48 to 4.76 (M = 4.35, SD = 0.32). The mean clarity ratings for the individual items ranged from 3.54 (Item 16) to 4.77 (Item 1). For the usefulness ratings, the overall mean ratings ranged from 3.33 to 4.19 (M = 3.93, SD = 0.28). The mean usefulness ratings for the individual items ranged from 2.31 (Item 21) to 4.77 (Item 1). Results of the clarity and usefulness ratings are presented in Table 1.

Discussion

The results of Study 1 suggest the need to revise or drop items that do not correspond directly to any of the *DSM–IV* symptoms of major depressive disorder. The lowest relevancy mean ratings were on Item 3 (past failure), Item 6 (punishment feelings), and Item 21 (loss of interest in sex), indicating low correspondence between these items and the *DSM–IV* depressive symptoms. Item

[2] Of the 10 expert raters invited to participate, 7 completed and returned the study packets. In compliance with the informed-consent procedures, we list (in alphabetical order) only the following expert raters: David Dozois, James Griffin, Philip Kendall, Thomas Ollendick, and Jane L. Wong.

Table 1
Expert and Adolescent Ratings of the Beck Depression Inventory—II Items

	Expert ratings				Adolescent ratings			
	Relevancy		Specificity		Clarity		Usefulness	
Item	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1. Sadness	4.57	1.13	4.00	1.26	4.77	0.44	4.77	0.44
2. Pessimism	4.29	0.76	4.00	0.89	4.00	0.82	4.38	0.51
3. Past failure	3.29	1.70	3.17	1.33	4.69	0.48	3.38	0.77
4. Loss of pleasure	4.71	0.49	4.00	0.89	4.62	0.77	4.46	0.78
5. Guilty feelings	4.29	0.95	3.67	0.82	4.15	0.55	3.69	0.48
6. Punishment feelings	3.43	0.79	3.00	0.89	4.62	0.77	3.38	0.55
7. Self-dislike	4.00	0.82	3.67	0.82	4.54	0.66	4.54	0.66
8. Self-criticalness	4.14	0.90	3.50	0.84	4.31	0.63	4.08	0.64
9. Suicidal thoughts or wishes	4.57	0.53	4.50	0.84	4.54	0.66	4.62	0.51
10. Crying	4.00	0.82	3.50	0.84	4.46	0.78	4.46	0.78
11. Agitation	4.14	1.46	2.67	1.51	4.69	0.48	3.08	0.76
12. Loss of interest	4.43	1.13	4.00	1.26	4.54	0.52	4.54	0.52
13. Indecisiveness	4.00	1.15	3.00	1.67	4.00	0.00	3.31	0.48
14. Worthlessness	4.71	0.49	4.17	1.17	4.69	0.48	4.69	0.48
15. Loss of energy	4.57	0.79	4.17	1.17	4.46	0.52	4.46	0.52
16. Changes in sleeping pattern	4.57	0.53	3.50	1.22	3.54	0.66	3.69	0.63
17. Irritability	4.29	1.70	3.17	1.47	4.08	0.76	3.23	0.83
18. Changes in appetite	4.00	1.29	3.33	1.37	3.62	0.87	3.77	0.44
19. Concentration difficulty	4.00	1.53	2.83	1.33	4.54	0.66	3.31	0.85
20. Tiredness or fatigue	4.29	0.95	3.33	1.37	4.31	0.85	4.31	0.85
21. Loss of interest in sex	2.71	0.76	2.50	0.84	4.38	0.77	2.31	0.75

11 (agitation), Item 19 (concentration difficulty), and Item 21 (loss of interest in sex) had the lowest specificity mean ratings. Consistent with study objectives, we did not drop or revise these items. Future studies should evaluate the clinical utility of the BDI–II severity scores by dropping or revising some of these items.

Study 2: Factor Structure of the BDI–II

Study 2 was designed to extend psychometric evaluation to the structure of the BDI–II in a sample of adolescent psychiatric inpatients. We conducted CFAs to examine the adequacy of fit of previously defined first-order solutions: (a) the two-factor solution reported in the BDI–II manual for the adult psychiatric outpatient samples, (b) the two-factor solutions reported in Dozois et al. (1998) for college undergraduates, and (c) the three-factor solution reported in Steer et al. (1998) for adolescent psychiatric outpatients. In addition, we assessed invariance of the two-factor solution reported in the manual across gender groups by using a multisample CFA procedure. Considering results of the CFAs, we decided to explore further the structure of the BDI–II. Furthermore, we report descriptive data and estimates of internal consistency of the BDI–II for the study sample.

Method

Participants

Participants were 205 boys and 203 girls recruited from the same child and adolescent units as those used in Study 1. The mean age of the combined sample was 15.09 years (SD = 1.39; range = 13–17 years). There were no significant differences between boys (M = 15.08 years, SD = 1.36) and girls (M = 15.09 years, SD = 1.41) in age, $t(406) = 0.04$, $p = .97$. The ethnic distribution of the sample included 83.8% Caucasian, 5.6% African American, 3.4% Hispanic/Latino American, 1.5% Asian American, and 5.6% mixed or other ethnicity. The principal diagnostic group distribution of the sample was as follows: 26.0% conduct disorder, 27.2% oppositional defiant disorder, 21.1% major depressive disorder, 10.0% adjustment disorder, 6.1% attention-deficit/hyperactivity disorder, and 9.6% other Axis I psychiatric disorders.

Procedure

All diagnoses were derived by the unit treatment teams composed of consultants, clinical, and nonclinical hospital staff. Multiple assessment methods were used, on all three units, to derive the final admission diagnoses. These included (a) *DSM–IV* structured diagnostic interviews conducted by the unit staff psychiatrists, (b) social-demographic interviews performed by the unit staff master's-level social workers, (c) structured psychological assessments conducted by the unit staff psychologists, (d) psychoeducational assessments performed by the school teachers, and (e) direct observation and monitoring of the patient by the professional and nonprofessional unit staff. The initial diagnosis was generally formulated by each unit staff psychiatrist, and the assessment information was subsequently used by the treatment team to form a consensus diagnosis. Augustine Osman (consultant for the agency) was present for a randomly selected number of cases on each unit to ensure consistency across the diagnostic processes. We did not evaluate the reliability and validity of these diagnoses. We excluded from participation adolescents who were not able to complete the questionnaires because of severe psychotic symptoms, organic impairment, or low intellectual functioning.

Parental or legal guardian consents were obtained during the intake sessions. We also obtained verbal permission from each potential participant before administering the questionnaire packets. Each participant individually completed a brief demographic information questionnaire and the BDI–II within 7 to 12 days of admission to the units. The research protocols were approved by the university Institutional Review Board and the hospital's program review committee.

Results

CFAs: Overview of the Analyses

We conducted the analyses in two parts using the BDI–II items. First (primary analyses), we used data from the combined sample to examine the adequacy of fit of the two- and three-factor models specified earlier. In addition, we evaluated an alternate one-factor model by constraining all 21 BDI–II items to load on a single factor. Second (secondary analyses), we evaluated the invariance of the two-factor oblique model reported in the manual across boys and girls by using a multisample estimation procedure. The BDI–II's model was used in the invariance analyses because it serves as the standard against which other models are frequently validated in the literature. We hypothesized that each model would provide a good fit to the present sample data.

We used the EQS for Windows 6.1 (Bentler & Wu, 2003) program because it provides multiple fit estimates including the Lagrange multiplier test (LMT) chi-square statistic (e.g., the associated probability value $\leq .05$ is generally used as evidence of nonvariance of an item) and the Satorra-Bentler scaled chi-square that allow for reliable and comprehensive evaluation of the adequacy of fit of models. The program also allows for the direct evaluation of multivariate normality in study sample data.

We analyzed all the models with the robust maximum-likelihood estimation procedure because preliminary analyses showed that the normalized Mardia's value (47.15) was high. We used contemporary cutoff values to test for the fit of the models. Specifically, we used consistency across all four robust fit indices as the initial criterion for evaluating further the stability of a model: (a) a normed fit index of .90 or higher, (b) a nonnormed fit index of .95 or higher, (c) a comparative fit index of .95 or higher, and (d) a root-mean-square error of approximation (RMSEA) value of .08 or less (see Browne & Cudeck, 1993; Hu & Bentler, 1999). Next, we examined (a) the R^2 estimates (i.e., values $\geq .20$ are supportive of less complex structure for the sample data) and (b) the standardized solutions (i.e., values $\geq .40$ are considered adequate) in determining the acceptability of a model.

In all the analyses, we set the variance of each factor to 1.0 and allowed the factors to be correlated (oblique) before testing the models. Items were constrained to load only on one factor; additionally, we did not allow for correlated errors. Results of the CFAs are shown in Table 2.

BDI–II two-factor oblique model. This oblique model consisted of the Somatic–Affective factor (Items 4, 10–13, and 15–21) and the Cognitive factor (Items 1–3, 5–9, and 14). The model met one of the four initial criteria for adequacy of fit. Examination of the R^2 values showed that the estimate for Item 21 (.064) was less than .20. Item 21 also loaded (.254) less than .40 on the Somatic–Affective factor. The correlation between these factors was high ($r = .92$). The model did not meet the preestablished fit criteria.

Dozois et al.'s (1998) two-factor oblique model. The two-factor model was composed of the Cognitive–Affective factor (Items 1–3, 5–9, 13, and 14) and the Somatic–Vegetative factor (Items 4, 10–12, 15–18, and 19–21). We found that this model also met only one of the four preestablished criteria. Similarly, Item 21 had a low R^2 value (.065), and it loaded .256 on the Somatic–Vegetative factor. The correlation between the factors was high ($r = .92$). This model did not provide acceptable fit to the sample data.

Steer et al.'s (1998) three-factor oblique model. This model consisted of the Cognitive factor (Items 2, 3, 7–9, 13, 14, and 19), the Somatic–Affective factor (Items 1, 4, 12, 15–18, 20, and 21), and the Guilt–Punishment factor (Items 5, 6, and 10). It met only one of the four initial criteria for adequacy of fit. Two of the items had R^2 values less than .20: Item 16 ($R^2 = .196$) and Item 21 ($R^2 = .063$). The correlations among the factors ranged from .84 to .94. This model provided inadequate fit to the sample data.

Alternate one-factor model. This model required that we constrain all 21 items to load on a single factor. The model met only one of the four initial criteria for adequacy of fit. Item 16 ($R^2 = .186$) and Item 21 ($R^2 = .054$) had low R^2 values, and Item 21 loaded (.23) less than the expected .40 preestablished criterion. This model provided poor fit to the present sample data.

None of the models tested met all the preestablished initial and final adequacy-of-fit criteria. Thus, our hypothesis was not supported. Because statistical comparisons of models are generally undertaken when an acceptable model has been established, we did not compare any of these models.

Analyses of Invariance: Overview

First, we established baseline models for each gender group. Next, we followed these steps in evaluating invariance across the gender groups. We tested for (a) group variant with no constraints imposed on any of the variables, (b) invariance of factor loadings, (c) invariance of factor correlations, and (d) invariance of factor loadings and factor intercorrelations across the gender groups.

In addition to the previous standard fit estimates, we used the LMT statistic and the standardized residuals to guide us in eval-

Table 2
Confirmatory Factor Analyses of the Beck Depression Inventory—II (BDI–II)

Model and study	S-B χ^2	*df*	*p*	NFI	NNFI	CFI	RMSEA (90% CI)
Standard fit estimates[a]			*ns*	.900	.950	.950	$\leq$.080
BDI–II manual (Beck et al., 1996)	326.16	188	<.01	.885	.941	.948	*.042* (.035, .050)
Dozois et al. (1998)	323.52	188	<.01	.886	.943	.949	*.042* (.034, .050)
Steer et al. (1998)	345.08	185	<.01	.879	.931	.939	*.046* (.038, .054)
Alternate model (present study)	371.83	189	<.01	.869	.923	.931	*.049* (.041, .056)

Note. Estimates meeting the preestablished cutoff scores are in italics. The total sample ($N = 408$) of 205 boys and 203 girls were included in all analyses. S-B χ^2 = Satorra–Bentler chi-square; NFI = normed fit index; NNFI = nonnormed fit index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; CI = confidence interval.

[a] Fit estimates recently recommended in the literature.

uating the adequacy of fit of each model. The lack of invariance of the BDI–II manual's two-factor model argues against the generalizability of this model across gender groups in this study.

Results of all the multigroup model analyses are presented in Table 3. The best baseline model for the boys (M1) required Item 13 to load on Factor 2 (Cognitive) to attain reasonable fit to the sample data. For the girls, Item 1 was allowed to load on Factor 1 (Somatic–Affective) to attain a reasonable fit of the baseline model (M2).

Next, when the parameters were allowed to differ for boys and girls (M3; total variant), the model approached a reasonable fit to the sample data. In evaluating invariance of factor loadings across gender groups (M4), we obtained evidence of noninvariance due to two items, Item 8 (LMT $\chi^2 = 5.08$, $p < .02$) and Item 18 (LMT $\chi^2 = 8.16$, $p < .01$). However, in the analysis involving invariance of factor intercorrelations (M5) across gender groups, we observed evidence of group invariance (LMT $\chi^2 = .87$, $p = .35$).

The results of the invariance of factor loadings and factor intercorrelations (M6) indicated evidence of noninvariance. One additional item, Item 7 (LMT $\chi^2 = 3.62$, $p < .05$) was not equivalent across boys and girls. Thus, although the two-factor structure was invariant across gender groups, evidence for similar item-factor compositions across gender groups could not be supported.

Exploratory Factor Analyses (EFAs) of the BDI–II Items: Overview of the Analyses

We conducted EFAs to explore alternate solutions of the BDI–II items. Although several EFA investigations with the different versions of the BDI have identified two-, three-, or four-factor solutions among adolescents (e.g., Bennett et al., 1997), no study to date has replicated clearly the item-factor solutions reported in the BDI manuals. Some clinical and nonclinical researchers (e.g., Dozois et al., 1998) have pointed to several methodological factors that may account for this problem: the different types of extraction techniques, rotation methods, sample sizes, and rules guiding the number of components to extract in the different studies. Our decision to explore further the factor structure of the BDI–II in the present sample was supported by the results from the current CFAs.

We used the maximum-likelihood parameter estimation (with robust standard errors and a mean adjusted chi-square test statistic procedure in the Mplus Version 2.12 (Muthén & Muthén, 2002) program to evaluate each component. Factors extracted were rotated using the promax procedure because we expected moderate to high correlations among potential factors from the same inventory. To address some prior measurement issues, we included two of the most accurate methods for determining the appropriate number of factors for extraction, the minimum average partial (MAP) and the parallel analysis (PA; 95th-percentile eigenvalues; see Zwick & Velicer, 1986); other extraction rules examined were the traditional scree plot and the eigenvalue greater than or equal to 1. Because the BDI–II and the *DSM–IV* propose that the structure of depressive symptoms in adults is similar to depressive symptom structures seen in adolescents, we expected all the BDI–II items to load meaningfully (i.e., .35 or higher) on the derived factor structures.

Following the recommendations of L. Muthén (personal communication, January 8, 2002), we looked initially at several factor retention rules, including (a) low root-mean-square residual and RMSEA indices (i.e., values ≤ .05), (b) a relative chi-square ratio (i.e., χ^2/df values ≤ 2.0), (c) residual variances (i.e., no high negative values), and (d) acceptable factor loadings (i.e., values ≥ .35). The final decision to retain a factor solution, however, was based on parsimony and interpretability of the factors. Results of the EFA are presented in Table 4.

EFA for the combined sample. The eigenvalue greater than or equal to 1, the scree plot, and the MAP rules suggested the extraction of a three-factor solution; the PA rule suggested the extraction of a one-factor solution. We extracted one to three factors and examined the solutions. Although the two- and three-factor solutions had comparable fit estimates, three items (Items 6, 10, and 21) failed to load adequately on any of the three factor solutions. Thus, we retained the two-factor solution because of parsimony and interpretability of the factors.

Factor 1 (eigenvalue = 8.80; % variance = 29.29) contained all 9 of the original Cognitive factor items reported in the BDI–II

Table 3
Multigroup Model Fit Estimates

Model	S-B χ^2	*df*	NFI	NNFI	CFI	RMSEA (90% CI)
Baseline models						
M1. Boys	220.05	188	.812	.963	.967	.029 (.000, .044)
M2. Girls	295.74	188	.842	.927	.935	.053 (.041, .064)
Invariance analyses						
M3. No constraints imposed	512.70	376	.828	.940	.946	.030 (.023, .036)
M4. Factor loading invariance	540.10	395	.818	.940	.943	.030 (.024, .036)
M5. Factor correlation invariance	513.53	377	.827	.940	.947	.030 (.023, .036)
M6. Factor loading and correlation invariance	530.48	394	.822	.943	.947	.029 (.022, .035)

Note. The total sample ($N = 408$) of 205 boys and 203 girls were included in all analyses. S-B χ^2 = Satorra–Bentler chi-square; NFI = normed fit index; NNFI = nonnormed fit index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; CI = confidence interval; M = model.

Table 4
Factor Loadings and Intercorrelations for the Beck Depression Inventory—II in Adolescent Psychiatric Inpatients

Item no. or factor	Combined samples		Boys		Girls	
	F1	F2	F1	F2	F1	F2
	Factor loadings					
1	*.40*	.33	*.49*	.14	*.50*	.28
2	*.58*	.12	*.55*	.20	*.78*	−.09
3	*.67*	.04	*.55*	.15	*.70*	.03
4	*.37*	.32	*.36*	.29	*.41*	.30
5	*.56*	.07	*.62*	−.09	*.48*	.21
6	*.40*	.07	.30	.09	*.44*	.09
7	*.73*	.06	*.68*	.06	*.70*	.11
8	*.80*	−.01	*.75*	.01	*.69*	.12
9	*.64*	.00	*.62*	.06	*.74*	−.12
10	.32	.19	*.40*	−.02	.33	.23
11	.20	*.43*	.30	*.35*	.11	*.53*
12	*.44*	.29	.27	*.46*	*.66*	.06
13	*.43*	.30	*.47*	.29	.29	*.46*
14	*.69*	.14	*.69*	.17	*.77*	.04
15	.26	*.44*	.14	*.56*	*.58*	.14
16	−.01	*.50*	.02	*.46*	.08	*.41*
17	*.37*	*.38*	*.44*	.23	*.50*	.28
18	.09	*.50*	−.14	*.54*	.27	*.45*
19	.09	*.67*	.02	*.72*	.29	*.50*
20	−.04	*.78*	.11	*.61*	−.04	*.86*
21	−.09	*.36*	.01	.32	−.14	*.37*
	Factor intercorrelations					
F1	—		—		—	
F2	.73	—	.65	—	.70	—

Note. Items loadings greater than or equal to .35 are in italics. F1 = Cognitive–Affective factor; F2 = Somatic factor.

manual for the adult psychiatric outpatients. Three affective items and 1 cognitive depressive item (Items 4, 12, 13, and 17) also loaded adequately on this factor; thus, this factor was named Cognitive–Affective. The second factor (eigenvalue = 1.15; % variance = 12.80) contained 8 items that were similar to the original 12 Somatic–Affective factor items; 1 of the items (Item 17) cross-loaded on Factor 1. Only 1 item (Item 10) failed to load adequately on any of the factors. This factor was named Somatic.

EFA for the boys. Examination of the extraction rules suggested that one (95% PA and MAP) to four (scree plot, and eigenvalue ≥ 1) factors could be extracted. We extracted one to four factors and examined the solutions. The chi-square values for the two-, three-, and four-factor solutions were nonsignificant (all *p*s > .05), suggesting good fit of each solution for the present sample data. However, the chi-square for the four-factor solution was significantly lower when compared with the chi-square values for the two-factor and three-factor structures. Examination of the item-factor loadings showed that only one item (Item 8, self-criticalness) loaded (.81) on the fourth factor; thus, this factor was dropped. Because the chi-square difference test showed no significant differences between the fit of the two-factor and three-factor structures ($p = .08$), we retained the two-factor solution because of parsimony and interpretability of the solutions.

Factor 1 (eigenvalue = 7.79; % variance = 24.51) was composed of 8 of the 9 original Cognitive factor items. In addition, 3 affective items and 1 cognitive item (Items 4, 10, 13, and 17) loaded adequately on this factor. The factor was named Cognitive–Affective. Factor 2 (eigenvalue = 1.37; % variance = 13.42) was composed of 7 of the 12 original Somatic–Affective factor items. Two items (Items 6 and 21) did not load adequately on any of the two factors. This factor was named Somatic.

EFA for the girls. The extraction rules suggested that one- (MAP and PA) to three- (eigenvalue ≥ 1 and scree plot) factor solutions could be extracted. We extracted one to three oblique factors and examined the solutions. The chi-square values for both the two- and three-factor solutions were statistically significant (all *p*s < .01). Additionally, the relative chi-square ratio and the RMSEA and root-mean-square residual indexes for both solutions were comparable. Examination of the solutions showed that two items (Items 4 and 6) failed to load adequately on any of the three-factor solutions. Thus, in terms of parsimony and interpretability, we retained the two-factor oblique solution.

Factor 1 (eigenvalues = 9.35; % variance = 34.65) contained nine items that were similar to the original nine Cognitive factor items in the manual. Four somatic items (Items 4, 12, 15, and 17) also loaded adequately on this factor. This factor was named Cognitive–Affective. Factor 2 (eigenvalue = 1.24; % variance = 10.67) was defined by seven of the original Somatic-–Affective factor items; it was named Somatic. Only one item (Item 10) failed to load adequately on any of the two factors.

Reliability Analyses

Cronbach alpha (Cronbach, 1951) estimates were computed to establish internal consistency of the BDI–II total and derived factor scales. The alpha estimates were evaluated for adequacy according to the guidelines provided by Cicchetti (1994). The mean interitem correlations were also computed to provide additional support for internal consistency.

For the total sample of 408 participants, the alpha estimate of the BDI–II total scale score was .93 (mean interitem correlation = .40). The internal consistency estimates of the Cognitive–Affective (α = .90; mean interitem correlation = .43) and the Somatic (α = 80; mean interitem correlation = .34) subscale scores were adequate. For the boys, estimates of internal consistency of the BDI–II total scale (α = .91; mean interitem correlation = .32) score as well as the Cognitive–Affective (α = .88; mean interitem correlation = .38) and the Somatic (α = .78; mean interitem correlation = .32) subscale scores were good. For the girls, we also obtained good internal consistency estimates for the BDI–II total score (α = .93; mean interitem correlation = .40) and for both the Cognitive–Affective (α = .92; mean interitem correlation = .46) and the Somatic (α = .79; mean interitem correlation = .35) subscale scores.

Gender Differences on the BDI–II

An independent-samples *t*-test analysis showed that the girls (*M* = 20.49, *SD* = 13.50) obtained significantly higher mean BDI–II total scores than did the boys (*M* = 15.02, *SD* = 11.07), $t(406) = 4.78$, $p < .01$, Cohen's $d = .48$ (small effect size). The mean BDI–II total score for the girls was 5.47 points higher than the total score for the boys. Based on the BDI–II screening criteria (Beck et al., 1996), the responses of the boys were in the mild range (14 to 19), and the responses of the girls were in the moderate range (20 to 28). The mean BDI–II total score for the total sample was 17.75 ($SD = 12.62$).

Study 3

The present study included only adolescents age 14 years and older because the psychometric characteristics of most of the concurrent validation instruments have not been established for youths age 13 and younger.

Method

Participants

A subset of the participants (161 boys and 158 girls, 14 to 17 years old) in Study 2 volunteered to complete the concurrent validation measures. The data from the consecutive admissions were combined because there were no significant differences between youths on demographic variables of age, gender composition, ethnicity, and diagnostic groups (all ps $> .05$; chi-square analyses). For the combined participants, the mean age was 15.67 years ($SD = 0.95$). The mean age of 15.65 years ($SD = 0.92$) for the boys, and the mean age of 15.68 years ($SD = 0.98$) for the girls did not differ significantly, $t(317) = 0.29$, $p = .77$. The ethnic distribution of the sample included 84.6% Caucasian, 5.0% African American, 3.5% Hispanic/Latino American, 1.3% Asian American, and 5.6% mixed or other ethnicity. The mean length of stay of the sample was approximately 46 days (range = 14–220 days). The mean Shipley Institute of Living Scale (Zachary, 1986) IQ of the sample was 103.29 ($SD = 9.28$), suggesting normal intellectual levels of functioning.

Procedure

As in Study 2, all admission diagnoses were derived by the multidisciplinary treatment team. The principal diagnostic group distribution of the sample was as follows: 33.2% conduct disorder, 22.6% oppositional defiant disorder, 22.6% major depressive disorder, 9.7% adjustment disorder, and 11.9% other Axis I psychiatric disorders (e.g., bipolar disorder and attention-deficit/hyperactivity disorder). Approximately 55.2% ($n = 176$) of the study participants had secondary diagnoses on Axis I. For additional descriptive information, we developed a mood disorder group ($n = 84$) that included individuals with major depressive disorder, dysthymic disorder, and bipolar disorder. We did not recruit adolescents with severe psychosis, organic impairment, or mental retardation.

Because provisional diagnoses were generally assigned when two or more of the core clinical staff members (specifically, the team psychiatrist and social workers) were not present for the full diagnostic discussion processes, data were not included for seven such cases. No formal reliability assessment data were obtained. Although we did not collect data on medication regimens, the medical records of the participants were examined in part to assess consistency between the assigned primary diagnosis and the treatment goals.

Parental or legal guardian consents were obtained during the intake sessions. We also obtained verbal permission from each potential participant before administering the questionnaire packets. The research protocol was approved by the university Institutional Review Board and the hospital's program review committee. All data were collected within 7 to 12 days of admission to the unit.

Measures

All participants completed the BDI–II, a brief demographic information questionnaire, and the following self-report measures of suicidal behavior, anxiety, hopelessness, and adolescent psychopathology.

Suicidal Behaviors Questionnaire—Revised. The Suicidal Behaviors Questionnaire–Revised (SBQ–R; Osman et al., 2001) is a brief four-item self-report measure of suicide-related behaviors. The SBQ–R Item 1 assesses prior suicide ideation and attempts, Item 2 taps frequency of suicide ideation, Item 3 taps threats of suicide, and Item 4 evaluates self-reported suicide likelihood. In clinical and nonclinical populations, reliability estimates for the SBQ–R range from .76 to .88. The total SBQ–R score (obtained by summing the ratings on all four items) of 8 (cutoff) has been identified as having adequate sensitivity (.87), specificity (.93), positive predictive value (.90), and negative predictive value (.99) in adolescent psychiatric inpatient samples (Osman et al., 2001). The total SBQ–R score has shown good evidence of concurrent and divergent validities in the assessment of adolescents and young adults and has been used as a criterion measure (e.g., see Gutierrez, Osman, Barrios, & Kopper, 2001; Osman, Barrios, et al., 2002). The SBQ–R total score was used as a measure of suicide-related behavior (risk) in evaluating evidence of concurrent validity for the BDI–II. The alpha estimates of the SBQ–R were .84, .84, and .85 for the boys, girls, and total sample, respectively.

Beck Anxiety Inventory. The Beck Anxiety Inventory (BAI; Beck & Steer, 1990) is a 21-item self-report measure that assesses the severity of anxiety symptoms, including subjective/panic and somatic. Each item is rated on a 4-point scale ranging from 0 (*not at all*) to 3 (*severely, I could barely stand it*). The total BAI score is obtained by summing the ratings on each item; the total score ranges from 0 to 63. This scale has excellent reliability and concurrent validity in adolescent populations (e.g., Jolly, Aruffo, Wherry, & Livingston, 1993; Osman, Hoffman, et al., 2002). Because scores on the BAI frequently correlate moderately to highly (values of .50 or higher) with scores on the Beck depression instruments (see Beck & Steer, 1993), scores on the BAI were used in part as covariates in examining relationships of the BDI–II with other validational instruments. In this study, the BAI total score was used as a measure of anxiety severity. The alpha estimates of the BAI total score for the boys (.92), girls (.91), and total sample (.92) were acceptable.

Beck Hopelessness Scale. The Beck Hopelessness Scale (BHS; Beck, Weissman, Lester, & Trexler, 1974) is a 20-item true–false self-report measure that is designed to assess negative attitudes about the future. It is a widely used instrument with clinical and nonclinical samples with good reliability and validity estimates (Gutierrez, Osman, Kopper, & Barrios, 2000; Lyndall, 2001). The BHS total score ranges from 0 to 20, with higher scores suggesting greater levels of negative attitudes about the future. The Kuder-Richardson formula estimates of the total BHS score in the boys (.91), girls (.93) and combined samples (.92) were adequate. The BHS total score was used, as a measure of suicide risk and hopelessness to establish concurrent validity for the BDI–II.

Minnesota Multiphasic Personality Inventory—Adolescents. The Minnesota Multiphasic Personality Inventory—Adolescents (MMPI–A; Butcher et al., 1992) is a well-validated, frequently used, self-report measure of adolescent psychopathology. It is composed of 478 items with a true–false response format. This inventory can be scored for a range of research and clinical scales, including validity (e.g., defensiveness), basic clinical (e.g., depression), and content (conduct problems). The manual reported excellent reliability estimates and strong evidence of convergent and discriminant validities.

Convergent, discriminant, and incremental validity estimates of selected MMPI–A content scales have been tested with psychiatric inpatients (Arita & Baer, 1998; McGrath, Pogge, & Stokes, 2002). We used two of the internalizing (Anxiety and Depression) and two of the externalizing (Anger and Conduct Problems) scales. In addition, we included one internalizing (Self-Esteem) and one externalizing (School Problems) scale to examine evidence of convergent and discriminant validity of the BDI–II. As in previous investigations, only valid MMPI–A protocols (e.g., dropping protocols with Lie Scale scores > 70; Infrequency scores > 90, and Defensiveness Scale scores > 70; see Archer, 1992; Butcher et al., 1992) were used in this study.

Results

Diagnostic Group (Means and Standard Deviations) Comparison

The trimmed means, Winsorized standard deviations, and reliability estimates of the BDI–II by diagnostic groups are presented in Table 5. We conducted trimmed-means (5%) *t*-test comparisons to evaluate the ability of scores on the BDI–II to discriminate between the responses of adolescents with major depressive disorder diagnoses and those with disorders of conduct: oppositional defiant and conduct disorder. The adjustment disorder, mood disorder, and other disorders groups were not included in these comparisons because of extensive overlap in symptoms with major depressive disorder. We predicted that youths with major depressive disorder would report higher depression severity symptoms than would youths with the externalizing disorders of conduct.

Results of the 5% trimmed-means *t*-test comparisons of the scores showed that the BDI–II scores for the major depressive disorder group was significantly higher than those obtained for the conduct disorder, $t(176) = –3.05$, $p < .01$, Cohen's $d = .48$, and the oppositional defiant disorder, $t(142)$ –3.88, $p < .01$, Cohen's $d = .67$ groups. Thus, our hypothesis received strong support.

Relationships of the BDI–II With Suicide Risk Measures

Pearson correlational analyses were conducted to examine the unique relationships of the BDI–II with the suicide risk measures. Recall that the BDI–II manual suggests that clinicians and researchers conduct assessment of suicide ideation when screening for depression severity.

Zero-order and partial correlations were initially computed between the BDI–II and related suicide risk measures, the SBQ–R, and the BHS scale scores. We expected that the BDI–II scores would be related moderately and significantly with scores on these risk measures (see Table 6). For the boys, the 21-item BDI–II total score correlated highly (values of .50 or higher) and significantly with scores on the BHS ($r = .62$), and the 20-item BDI–II also correlated highly with the SBQ–R ($r = .51$).[3] When the BAI scores were partialed out, these relationships remained statistically significant (BDI–II score vs. SBQ–R score, $r = .43$; BDI–II score vs. the BHS score, $r = .59$).

Similarly for the girls, the correlation between the 21-item BDI–II total scores and scores on the BHS ($r = .69$) was high and statistically significant, and the correlation between the 20-item BDI–II total scores and scores on the SBQ–R ($r = .60$) was high and significant. When the BAI total scores were controlled statistically, these correlations also remained statistically significant (BDI–II vs. SBQ–R, $r = .44$, $p < .01$; BDI–II vs. BHS, $r = .58$).

Independent correlational analyses showed that the magnitude of the correlations between the BDI–II and each risk measure was comparable for both boys and girls (all *p*s > .05, two-tailed tests). Our hypothesis was supported in these initial analyses; controlling for anxiety-related symptoms (BAI total scores) did not substantially change the magnitude of these relationships.

Relationships of BDI–II With Selected Measures of Adolescent Psychopathology

Table 7 shows the zero-order and partial correlations between the BDI–II total scores and the selected MMPI–A content scale scores for the total sample. Preliminary zero-order correlations showed high and significant relationships between the BDI–II total scores and the BAI total scores ($r = .61$, $p < .01$). Thus, the BAI total scores were controlled statistically (partial correlation) in each analysis. In addition, a Bonferroni correction (.05/6) estimate was used with the corrected *p* value set at .008 to minimize a Type I error. To examine evidence of convergent and discriminant validity, we hypothesized that scores on the BDI–II would correlate higher with the internalizing scale scores (Anxiety, Depression, and Low Self-esteem) than with the externalizing scale scores (Anger, Conduct Problems, and School Problems).

The BDI–II scores correlated moderately (values = .40–.60) with scores on the internalizing measures and low (values = .20–.39) with scores on the externalizing measures. When we controlled statistically for the BAI total scores, only one of these correlations was nonsignificant. Results of the dependent correlation analyses (Steiger's *z* test) showed that the correlation between the BDI–II and the convergent measures were significantly higher than those obtained between the BDI–II total scores and the discriminant validity measures. Thus, our hypothesis also received strong support.

General Discussion

The present investigation is the first to undertake an extensive examination of the content validity and a number of other psychometric properties of the BDI–II. Previous investigations have not addressed adequately most of these psychometric issues with adolescent inpatients. Given the strong relationships between depressive symptoms and other psychiatric disorders such as anxiety and adjustment disorders (e.g., see Brady & Kendall, 1992; Ollendick, Seligman, & Butcher, 1999), the findings of the present study may have clinical relevance in the child and adolescent depression literature. Study 1 was carried out to examine evidence of content validity. Indeed, the content validity of any assessment instrument, like any form of validity cannot be assumed, it must be established empirically. Ratings by experts and psychiatric inpatients offered some initial directives for making the BDI–II more suitable for use with adolescents.

Specifically, the expert raters assigned low ratings (i.e., values below the overall mean ratings) of relevancy to several items. Some of these items were as follows: Item 3 (past failure), Item 6 (punishment feelings), and Item 21 (loss of interest in sex). It should also be noted that the BDI–II does not contain an equal number of items for each of the *DSM–IV* depressive symptoms. For example, the *DSM–IV* "diminished interest or pleasure" symptom corresponds to three BDI–II items (Items 4, 12, and 21); most of the other *DSM–IV* items correspond to either one or two BDI–II items. Future studies should evaluate the clinical utility of an unequal number of BDI–II items for each of the *DSM–IV* symptoms in screening for depression severity. Overall, the expert raters rated the total BDI–II score as being highly useful in screening for depression severity with adolescents. Additional comments provided by these expert raters (see *Results* of Study 1) might be useful to developers of the BDI–II in future revisions of this instrument.

[3] We dropped BDI–II Item 9 (suicidal thoughts or wishes) in the comparisons involving the SBQ–R scores.

Table 5
Trimmed Means, Winsorized Standard Deviations, and Internal Consistency Reliability Estimates

	BDI–II				Reliability estimates of the BDI–II	
Diagnostic group[a]	*n*	*M*	*SD* (Winsorized)	95% CI	α	Mean interitem *r*
Major depressive disorder	72	21.95	14.55	19.07, 25.88	.94	.43
Conduct disorder	106	15.99	10.66	14.47, 18.93	.91	.34
Oppositional disorder	72	13.38	10.71	11.62, 16.79	.92	.35
Adjustment disorder[b]	31	17.20	11.21	13.45, 21.71	.92	.34
Mood disorders[b,c]	84	22.48	14.53	19.78, 26.07	.94	.42
Other disorders[b]	38	14.20	14.03	10.71, 19.55	.95	.46

Note. BDI–II = Beck Depression Inventory—II; CI = confidence interval.
[a] Scores for the mood and major depressive disorder groups were in the moderate range (20 to 28); scores for all other groups were in the mild range (14 to 19). [b] These groups were not included in the group comparison analyses in the text. [c] Includes major depressive, bipolar, and dysthymic disorders.

Results of the clarity ratings of the BDI–II items by the inpatients indicated that most of the items were easy to read and understand; only two items were seen as very difficult to read: Items 16 (changes in sleeping pattern) and 18 (changes in appetite). We should note that both of these items have multiple response options that might be difficult for youths with severe disturbances in mood to complete. Specifically, each requires youngsters to attend carefully to all seven options before completing the ratings.

Study 2 was designed to examine the factor structure of the BDI–II in a sample of adolescent psychiatric inpatients. Unlike previous factor analytic investigations with the BDI–II, we followed a variety of contemporary psychometric recommendations (e.g., see Floyd & Widaman, 1995; Zwick & Velicer, 1986). For example, we examined multiple evaluation criteria in determining the appropriate number of factors to extract and retain.

Results of the CFAs showed that none of the models tested met the preestablished criteria for use with the present sample data. Similar results were obtained in the invariance analyses. Thus, we conducted EFAs to identify specific BDI–II factor structures for the sample data. We examined a range of solutions for adopting the most interpretable and parsimonious solution for retention.

For the boys, girls, and combined samples, we retained the two-factor oblique solution. The BDI–II manual also reported a two-factor solution for the adult inpatients. However, we observed substantial differences between the solutions for the adolescents and those reported in the manual. Specifically, in contrast with the solutions for the adult inpatients, the first factor we extracted was characterized by both cognitive and affective items. Thus, for these adolescents, mixed cognitive–affective symptoms of depression may be more important than the somatic symptoms.

Table 6
Relationships of the Beck Depression Inventory—II Scales With the Suicide Risk Measures

	Boys (*n* = 161)		Girls (*n* = 158)		
Measure	Zero-order *r*	Partial *r*	Zero-order *r*	Partial *r*	*z*[a]
BAI	.53**		.63**		1.33
BHS	.62**	.59**	.69**	.58**	1.09
SBQ–R[b]	.51**	.43**	.60**	.44**	1.15

Note. BAI = Beck Anxiety Inventory; BHS = Beck Hopelessness Scale; SBQ–R = Suicidal Behaviors Questionnaire–Revised.
[a] Independent correlation analyses; all *p*s > .05. [b] We removed scores on Item 9 (suicidal thoughts or wishes) in computing these coefficients.
** $p \leq .01$.

Further examination of the item-factor compositions showed similarities for boys and girls in the reporting of depressive symptoms. For example, for both boys and girls, the predominant symptoms contained in Factor 1 (Cognitive–Affective) were cognitive; the predominant symptoms in Factor 2 were Somatic. Only minor differences were observed between the gender groups. For example, for the boys, two of the items (Items 6 and 21) failed to load high on any of the factors. For the girls, one item (Item 10) failed to load high on both factors, and the Somatic factor included one cognitive item (Item 13). Future studies should attempt to replicate the present findings to validate the structure of the BDI–II for boys and girls.

Additional purposes of Study 2 were to evaluate the internal consistency of the BDI–II and to report descriptive data for the present sample. Results of the reliability analyses were good across the subsamples including boys and girls. The alpha estimates of the BDI–II were comparable with those reported in most recent studies (e.g., see Beck et al., 1996; Krefetz et al., 2002; Kumar et al., 2002; Steer et al., 1998). The mean BDI–II total scores for the boys and girls were similar to those reported by Steer et al. for the adolescent psychiatric outpatients. Similar to the frequent reports in the empirical literature (e.g., Nolen-Hoeksema & Girgus, 1994), girls obtained significantly higher depression severity scores than boys in the present study. Results of the EFA, however, suggest similarities in the structure and compositions of depressive symptoms expressed by boys and girls.

Findings from Study 3 provided additional information regarding scores on the BDI–II. First, we provided descriptive data for the BDI–II across a variety of team-derived diagnostic groups. Given that the treatment goals for the study participants were based on these team-derived diagnoses (as commonly conducted in most traditional psychiatric inpatient settings), the data reported in the present study may serve as appropriate preliminary baselines in future investigations of the BDI–II with adolescents in most traditional psychiatric inpatient settings.

Table 7
Relationships of the Beck Depression Inventory—II With Selected MMPI–A Content Scales

Measure	Zero-order *r*	Partial *r*[a]	Dependent comparison	*z*
Internalizing measures				
Adol-Anxiety	.54**	.30**	Adol-Anxiety vs. Adol-Anger	4.72**
			Adol-Anxiety vs. Adol-Conduct Problems	3.91**
			Adol-Anxiety vs. Adol-School Problems	6.33**
Adol-Depression	.55**	.39**	Adol-Depression vs. Adol-Anger	4.72**
			Adol-Depression vs. Adol-Conduct Problems	4.26**
			Adol-Depression vs. Adol-School Problems	6.66**
Adol-Self Esteem	.47**	.30**	Adol-Self Esteem vs. Adol-Anger	2.93**
			Adol-Self Esteem vs. Adol-Conduct Problems	2.32*
			Adol-Self Esteem vs. Adol-School Problems	4.94**
Externalizing measures				
Adol-Anger	.32**	.14**		
Adol-Conduct Problems	.35**	.16**		
Adol-School Problems	.20**	.05		

Note. MMPI–A = Minnesota Multiphasic Personality Inventory—Adolescents. Adol = adolescent.
[a] Controlling for the Beck Anxiety Inventory total scores.
* $p \leq .05$. ** $p \leq .01$.

Second, the results of Study 3 showed that scores on the BDI–II are adequately related to measures of suicide risk, suggesting strong evidence for convergent validity. For both boys and girls, depression severity was significantly related to suicide ideation and hopelessness. Additional evidence of convergent and discriminant validity was observed in the pattern of the relationships between the BDI–II total scores and the selected MMPI–A scale scores. Specifically, the BDI–II total score correlated moderately (low and nonredundant) and significantly with measures designed to tap common constructs including anxiety, depression, and low self-esteem (e.g., see Clark & Watson, 1991), even after for the high comorbid anxiety symptoms were controlled for. The BDI–II scores correlated lower with measures designed to tap unrelated constructs including anger, conduct problems, and school problems, suggesting evidence of discriminant validity. These findings of convergent and discriminant validities were strengthened by results of the dependent correlational analyses; scores on the convergent measures were significantly related to BDI–II scores when compared with scores on the discriminant measures.

The present study has several limitations. First, as with most cross-sectional investigations, the findings do not imply direct causation or mediation. Future investigations that use longitudinal procedures could examine relationships among the study measures over time. Second, the study participants included mostly Caucasian inpatient youths. Future studies might include adolescents from diverse ethnic groups and other geographic settings. In addition, the inclusion of both nonclinical and outpatient participants might help enhance the generalizability of the findings. Third, the use of the *DSM–IV* as criteria for evaluating the content validity of the BDI–II does not suggest that scores on the BDI–II can be used to make a diagnosis of major depressive disorder. We note that although this scale contains symptoms that correspond to those in the *DSM–IV*, it does not specify any criteria for use as a diagnostic tool. Fourth, we did not ask the expert raters or adolescents to generate potential items that could be included in the revision of the BDI–II. Fifth, we did not attempt to determine specific cutoff scores, as in previous investigations (see the BDI–II manual), for determining depression severity. Despite these limitations, the present study is the first to present useful clinical data addressing some of the limitations in the BDI–II research with adolescent psychiatric inpatient samples. The data suggest that the BDI–II is a promising measure of depression severity for use with adolescent psychiatric inpatients. We encourage replications of the present findings as well as validation of a modified set of BDI–II items with high mean relevancy and clarity ratings.

References

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

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Archer, R. P. (1992). *MMPI–A*: *Assessing adolescent psychopathology.* Hillsdale, NJ: Erlbaum.

Arita, A. A., & Baer, R. A. (1998). Validity of selected MMPI–A content scales. *Psychological Assessment, 10,* 59–63.

Arnau, R. C., Meagher, M. W., Norris, M. P., & Bramson, R. (2001). Psychometric evaluation of the Beck Depression Inventory—II with primary care medical patients. *Health Psychology, 20,* 112–119.

Beck, A. T., & Steer, R. A. (1990). *Manual for the Beck Anxiety Inventory.* San Antonio, TX: Psychological Corporation.

Beck, A. T., & Steer, R. A. (1993). *Manual for the revised Beck Depression Inventory.* San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory—II.* San Antonio, TX: Psychological Corporation.

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

Beck, A. T., Weissman, A., Lester, D., & Trexler, M. (1974). The measurement of pessimism: The Hopelessness Scale. *Journal of Consulting and Clinical Psychology, 42,* 861–865.

Bennett, D. S., Ambrosini, P. J., Bianchi, M., Barnett, D., Metz, C., & Rabinovich, H. (1997). Relationship of Beck Depression Inventory factors to depression among adolescents. *Journal of Affective Disorders, 45,* 127–134.

Bentler, P. M., & Wu, E. J. C. (2003). *EQS structural equations program* (Version 6.1) [Computer software]. Encino, CA: Multivariate Software.

Brady, E. U., & Kendall, P. C. (1992). Comorbidity of anxiety and depression in children and adolescents. *Psychological Bulletin, 111,* 244–255.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). *Manual for administration, scoring, and interpretation of the Minnesota Multiphasic Personality Inventory for Adolescents: MMPI–A.* Minneapolis: University of Minnesota Press.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290.

Clark, L. A., & Watson, D. (1991). A tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100,* 316–336.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–315.

Dozois, D. J. A., Dobson, K. S., & Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory—II. *Psychological Assessment, 10,* 83–89.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286–299.

Gutierrez, P. M., Osman, A., Barrios, F. X., & Kopper, B. A. (2001). Development and initial validation of the Self-Harm Behavior Questionnaire. *Journal of Personality Assessment, 77,* 475–490.

Gutierrez, P. M., Osman, A., Kopper, B. A., & Barrios, F. X. (2000). Why young people do not kill themselves: The Reasons for Living Inventory for Adolescents. *Journal of Clinical Child Psychology, 29,* 177–187.

Haynes, S. N., Richard, D., & Kubany, E. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Jolly, J. B., Aruffo, J. F., Wherry, J. N., & Livingston, R. (1993). The utility of the Beck Anxiety Inventory with inpatient adolescents. *Journal of Anxiety Disorders, 7,* 95–106.

Krefetz, D. G., Steer, R. A., Gulab, N. A., & Beck, A. T. (2002). Convergent validity of the Beck Depression Inventory—II with the Reynolds Adolescent Depression Scale in psychiatric inpatients. *Journal of Personality Assessment, 78,* 451–460.

Kumar, G., Steer, R. A., Teitelman, K. B., & Villacis, L. (2002). Effectiveness of the Beck Depression Inventory—II subscales in screening for major depressive disorders in adolescent psychiatric inpatients. *Assessment, 9,* 164–170.

Lyndall, S. (2001). Further validity and reliability evidence for Beck Hopelessness Scale scores in a nonclinical sample. *Educational and Psychological Measurement, 61,* 303–316.

McGrath, R. E., Pogge, D. L., & Stokes, J. M. (2002). Incremental validity of selected MMPI–A content scales in an inpatient setting. *Psychological Assessment, 14,* 401–409.

Muthén, L. K., & Muthén, B. O. (2002). *Mplus statistical analysis with latent variables: User's guide.* Los Angeles: Author.

Nolen-Hoeksema, S., & Girgus, J. S. (1994). The emergence of gender differences in depression during adolescence. *Psychological Bulletin, 115,* 424–443.

Ollendick, T. H., Seligman, L. D., & Butcher, A. T. (1999). Does anxiety mitigate the behavioral expression of severe conduct disorder in delinquent youths? *Journal of Anxiety Disorders, 13,* 565–574.

Osman, A., Bagge, C. L., Gutierrez, P. M., Konick, L. C., Kopper, B. A., & Barrios, F. X. (2001). The Suicidal Behaviors Questionnaire—Revised (SBQ–R): Validation with clinical and nonclinical samples. *Assessment, 8,* 443–454.

Osman, A., Barrios, F. X., Gutierrez, P. M., Wrangham, J. J., Kopper, B. A., Truelove, R. S., & Linden, S. C. (2002). The Positive and Negative Suicide Ideation (PANSI) Inventory: Psychometric evaluation with adolescent inpatient samples. *Journal of Personality Assessment, 79,* 512–530.

Osman, A., Downs, W. R., Barrios, F. X., Kopper, B. A., Gutierrez, P. M., & Chiros, C. E. (1997). Factor structure and psychometric characteristics of the Beck Depression Inventory—II. *Journal of Psychopathology and Behavioral Assessment, 19,* 359–376.

Osman, A., Hoffman, J., Barrios, F. X., Kopper, B. A., Breitenstein, J. L., & Hahn, S. K. (2002). Factor structure, reliability, and validity of the Beck Anxiety Inventory in adolescent psychiatric inpatients. *Journal of Clinical Psychology, 58,* 443–456.

Reynolds, W. M. (1987). *Reynolds Adolescent Depression Scale: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Steer, R. A., Ball, R., Ranieri, W. F., & Beck, A. T. (1999). Dimensions of the Beck Depression Inventory—II in clinically depressed outpatients. *Journal of Clinical Psychology, 55,* 117–128.

Steer, R. A., Kumar, G., Ranieri, W. F., & Beck, A. T. (1998). Use of the Beck Depression Inventory—II with adolescent psychiatric outpatients. *Journal of Psychopathology and Behavioral Assessment, 20,* 127–137.

Twenge, J. M., & Nolen-Hoeksema, S. (2002). Age, gender, race, socioeconomic status, and birth cohort differences on the Children's Depression Inventory: A meta-analysis. *Journal of Abnormal Psychology, 111,* 578–588.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children—Third Edition.* San Antonio, TX: Psychological Corporation.

Zachary, R. A. (1986). *Shipley Institute of Living Scale, revised manual*. Los Angeles: Western Psychological Services.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432–442.

(*Appendix follows*)

Appendix

Overall Beck Depression Inventory—II (BDI–II) Ratings

Please answer the following questions regarding the use of the BDI–II with adolescents, ages 13 to 17 years.

1. Two dimensions (factors) of the BDI–II items have been reported frequently in the clinical literature: somatic–affective and cognitive. Do these dimensions represent adequately *all* the dimensions of the major depressive disorder construct for adolescents?
   ____ Yes ____ No ____ Not Sure

2. Does the BDI–II contain all the *essential symptoms* seen in adolescents (ages 13 to 17 years) who are clinically diagnosed with major depressive disorder?
   ____ Yes ____ No ____ Not Sure

3. Please *list all* the items of the BDI–II that are considered *inappropriate* for use with adolescents, ages 13 to 17 years:
   a) ______________________________, d) ______________________________,
   b) ______________________________, e) ______________________________,
   c) ______________________________, f) ______________________________.

4. Overall, how useful is the BDI–II total score to the assessment of major depressive disorder symptoms for adolescents, ages 13 to 17 years?

1	2	3	4	5	6
Not at all useful					Extremely useful

Additional comments:

Thank you very much for your active participation in our research project.

Received April 15, 2002
Revision received December 30, 2003
Accepted January 6, 2004 ■