Journal of Clinical Psychology in Medical Settings, Vol. 12, No. 2, June 2005 (© 2005)
DOI: 10.1007/s10880-005-3273-x

Measurement of Depressive Symptoms in Women With Breast Cancer and Women With Clinical Depression: A Differential Item Functioning Analysis

Niels G. Waller,[1,4] Bruce E. Compas,[1] Steven D. Hollon,[3] and Ellen Beckjord[1,2]

Differential item functioning (DIF) analyses of the Beck Depression Inventory-II (BDI-II) were conducted on samples of 267 women with breast cancer and 294 women with clinical depression. Patterns of items in which there was significant and nonsignificant DIF were identified using statistical tests and measures of DIF effect size. At the most general level, 15 of 21 BDI-II items were associated with nontrivial DIF suggesting that the item responses of these samples do not reflect the same underlying construct. Factor analyses of the BDI-II using a psychometrically defensible method for item level factor analysis supported the conclusions from the DIF analyses. These findings suggest that researchers and practitioners should apply caution when interpreting self-report depression symptoms in breast cancer patients.

KEY WORDS: breast cancer; depression; measurement.

Accurate assessment of depression and depressive symptoms in medically ill patients is of paramount importance. Studies have shown that subclinical symptoms of depression and diagnoses of major depressive disorders are implicated in the psychological and physical health of patients with a wide range of conditions. This includes patients recovering from acute events such as myocardial infarction (e.g., Frasure-Smith, Lesperance, Juneau, Tlarijic, & Bourassa, 1999), and patients suffering from chronic conditions such as HIV infection and AIDS (Rosenberg et al., 2001) and cancer (McDaniel, Musselman, Porter, Reed, & Nemeroff, 1995). Depression is a possible consequence of the stress associated with some medical conditions, is a predictor of course and recovery, and may be an important source of comorbidity that can affect response to treatment.

In spite of its potential importance, the measurement of depression and depressive symptoms associated with disease has been hindered by several factors. Foremost among these problems is the possible confounding of symptoms of depression with symptoms of some diseases and with possible side effects of treatment. These concerns are particularly important in the assessment of depression in cancer patients, as some symptoms of some types of cancer and the side effects of adjuvant therapies can mimic symptoms of depression (Croyle & Rowland, 2003; Raison & Miller, 2003). For example, some forms of chemotherapy are associated with increased fatigue, loss of energy, decreased appetite, loss of sexual drive, and impairment in concentration and attention, all of which are focal symptoms of depression. Disentangling symptoms that are attributable to depression as contrasted with those that are due to disease or side effects of treatment are particularly difficult when depressive symptoms are measured with self-report questionnaires, such as the BDI-II (e.g., the Beck Depression Inventory-II; Beck, Steer, & Brown, 1996; see Dozois & Covin, in press, for a scholarly review). In both

[1]Department of Psychology and Human Development, Vanderbilt University, Nashville, Tennessee.
[2]Present address: Division of Cancer Prevention, National Cancer Institute, National Institute's of Health, Department of Health and Human Services, Bethesda, MD.
[3]Department of Psychology, Vanderbilt University, Nashville, Tennessee.
[4]Correspondence should be addressed to Niels Waller, Department of Psychology and Human Development, Peabody College, Vanderbilt University, Nashville, Tennessee 37203; e-mail: niels.waller@vanderbilt.edu.

research and clinical practice, self-report inventories are typically examined only in terms of their aggregate scores and little or no attention is given to other factors that could lead to endorsement of some symptoms as a result of disease or treatment processes.

Breast cancer patients are an optimal population in which to examine factors that may affect depressive symptoms. Breast cancer is the most commonly diagnosed cancer among women in the U.S. (American Cancer Society, 2003), and it has been the focus of extensive and intensive research on the psychological correlates of the disease and its treatment (Compas & Leucken, 2002). However, few studies (Ritterband & Spielberger, 2001) have directly compared the rates and patterns of endorsement of depressive symptoms among women with breast cancer, women free of disease, and women with clinical depression.

The few studies that have been conducted in this area (summarized in Spijker, Trijsburg, & Duivenvoorder, 1997) typically find that cancer patients produce elevated ratings on self-report depression scales when compared to healthy controls. Unfortunately, this finding is difficult to interpret because many researchers have not considered whether depression scales measure depression—and only depression—in cancer patients. Asking this logically prior question, Ritterband and Spielberger (2001) recently noted that "depression measures usually include items that assess somatic and performance difficulties that may not be just symptoms of depression, but rather consequences of disease treatment . . . [t]herefore, the total scores of cancer patients on measures of depression may overestimate the severity of depression, resulting in many false positive findings" (p. 86). These remarks point to the need for focusing our analyses on a structural level that is lower than that provided by scale total scores. The most obvious lower level considers item response data.

Potentially, item-level analyses can uncover population differences in scale structure and construct composition that are hidden from subscale or total scale analyses. This could occur, for instance, if the aggregate scores failed to measure similar dimensions in different populations. Ultimately, multivariate techniques such as factor analysis can elucidate scale and item dimensionality in samples that are homogenous with respect to diagnosis, but these analyses must be conducted with appropriate psychometric models. To our knowledge, no study has explored the BDI-II factor structure in cancer patients using psychometrically optimal methods for ordered categorical items (e.g., binary or Likert items, see Bock, Gibbons, & Muraki, 1988; Waller, 2003).

A second desiderata of item level analyses is the ability to explore trait-by-group interactions via models of differential item functioning (DIF; Camilli & Shepard, 1994; Holland & Wainer, 1993; Thissen, Steinberg, & Gerrard, 1986). Recent DIF analyses of the BDI (Kim, Pilkonis, Frank, Thase, & Reynolds, 2002; Santor, Ramsay, & Zuroff, 1994a) and other depression measures (Santor & Coyne, 2001; Santor, Zuroff, Cervantes, Palacios, & Ramsay, 1995) have advanced our understanding of depression assessment in various populations. Thus, a DIF analysis of the BDI-II in women with breast cancer is likely to enhance our understanding of depression assessment in this important population (American Cancer Society, 2003).

In the remainder of this paper we define DIF, describe how to measure DIF with standard statistical software, and then perform a DIF analysis of the BDI-II in women with breast cancer and women with clinical depression. We then explore the factor structure of the BDI-II using a psychometrically defensible method for item level factor analysis (Knol & Berger, 1991; Waller, 2003) and a rotation method (Schmid & Leiman, 1957) that simultaneously elucidates the general and group factors of the inventory. We conclude by discussing the implications of our results for depression assessment in women with breast cancer.

Measuring Differential Item Functioning With Logistic Regression

Stated plainly, a DIF study asks the following question: For trait levels, θ_i, does the probability (P) of an item response (U) for an individual from Group A differ from that of an individual from Group B? Notice in this definition that we are not comparing group means (e.g., by t-tests) or group item response rates (by χ^2 tests). On the contrary, all group comparisons are made with trait levels held constant. This is an important feature of DIF methodology that allows DIF studies to identify biased items.

An item is biased if it produces different item-trait regression functions in different groups. To understand this statement, consider the definition of an unbiased item. If an item is unbiased—that is, if an item shows no DIF—then the conditioned response probabilities can be expressed:

$$P_j(U|\theta_i, g = \mathrm{A}) = P_j(U|\theta_i, g = \mathrm{B}), \qquad (1)$$

where g is a group designator, j is an item index, and all other terms are defined as above. In plain English, this equation states that the probability that an individual

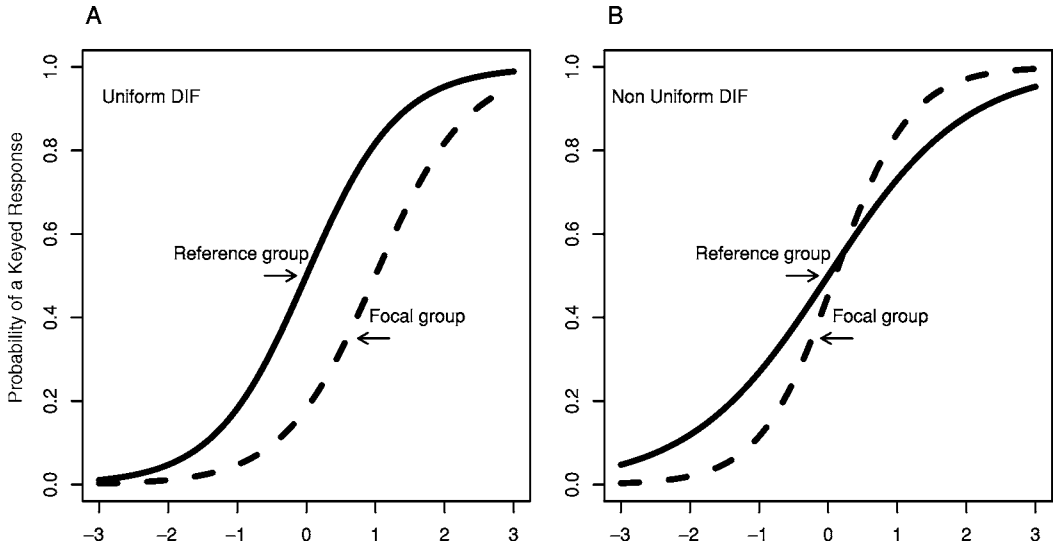

Fig. 1. Example item response functions showing uniform and nonuniform DIF.

from Group A with trait level, θ_i, responds to Item j in the keyed direction is equal to the probability that an individual from Group B with trait level, θ_i, responds to Item j in the keyed direction. At a conceptual level, DIF is present when these probabilities are statistically different.

Early DIF studies were conducted to identify aptitude and achievement items that showed bias against traditionally defined majority or minority groups (such as Whites, Hispanics, and Blacks). In more recent applications these group designators (i.e., majority and minority) are not always meaningful. Consequently, in two group designs, it is now customary to designate one group as the Reference (R) group and the other group as the Focal (F) group. In the present study the Reference group is composed of women with depression and the Focal group is composed of women with breast cancer. A DIF analysis allows us to determine whether the breast cancer and noncancer patients with similar levels of depression have different probabilities of endorsing items on the BDI-II.

Several methods are available for detecting DIF at the item and scale levels (Holland & Wainer, 1993; Millsap & Everson, 1993). The most popular methods are the Mantel-Haenszel test (Holland & Thayer, 1988), methods based on Item Response Theory (IRT; Raju, 1988; Shealy & Stout, 1993; Thissen, Steinberg, & Wainer, 1993), and methods based on logistic regression (Swaminathan & Rogers, 1990). These procedures have different attractive properties and differ among themselves in their ability to identify various forms of DIF. For instance, the Mantel-Haenszel test detects uniform DIF whereas the IRT and logistic regression methods identify both uniform and nonuniform DIF. These types of DIF are illustrated in Fig. 1.

Figure 1A displays two item trait regression lines for a dichotomously scored item. Each regression line corresponds to an item response function (IRF) for a specific group (i.e., the Reference or Focal group). Values along the x-axis denote trait scores whereas values along the y-axis denote item response probabilities (i.e., probabilities of endorsing the item in the keyed direction). Several features of this figure merit comment. First, notice that the item-trait regression lines are nonlinear. Linear item response functions would yield nonadmissable probability estimates for very low or very high trait scores (i.e., numbers that fall outside of the [0, 1] range); thus linear response functions are rarely used in item response models. Second, notice in the first figure that the IRFs for the Reference and Focal groups do not intersect or touch one another. If the two IRFs were identical, such that only one line was visible in the plot, then the item would show no bias and DIF would be absent. In this context, bias is defined as *group-specific* item endorsement probabilities for individuals with identical trait scores. From a psychometric perspective, item bias reveals itself as either uniform or nonuniform DIF. Uniform DIF is characterized by nonintersecting IRFs. Figure 1A shows an example of uniform DIF. Notice in this figure that, when trait level is held constant, Focal group members are less likely than Reference group members to endorse the item at all trait levels.

Nonuniform DIF is visually characterized by intersecting IRFs. Figure 1B displays an example of nonuniform DIF. Psychologically speaking, nonuniform DIF is particularly interesting because it suggests that relative item difficulty (or probability of symptom expression) changes as a function of both trait level and group membership. Statistically speaking, this is an instance of a group-by-trait interaction. Notice in Fig. 1B that low scoring members of the Reference group are more likely to endorse the item than are comparable members of the Focal group. At higher trait levels this pattern is reversed. In the context of depression assessment, findings of this type indicate that symptom expression is multidetermined and that group membership is a contributing factor.

As noted previously, there are several methods for detecting DIF in dichotomously scored items. Many of these methods require application-specific software (Thissen, 1991; Waller, 1998a, 1998b; Waller, Thompson, & Wenk, 2000). In the present study we used a method that is suitable in moderately sized samples (Rogers & Swaminathan, 1993; in general IRT and other latent variable DIF procedures require large samples to produce reliable findings) and can be tested with widely available software. Specifically, the procedure is based on a logistic regression model (Swaminathan & Rogers, 1990) and thus any statistics package that can run logistic regression can be used to test the following DIF model. When compared to alternative approaches, the logistic regression model is highly flexible (the model can accommodate covariates or additional conditioning variables), easy to implement, and statistically powerful (Clauser & Mazor, 1998; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Algebraically, the model can be expressed as follows:

$$P_j(U_j = 1|\theta, g) = \frac{e^{\tau_0+\tau_1\theta_i+\tau_2 g+\tau_3(\theta\times g)}}{1 + e^{\tau_0+\tau_1\theta+\tau_2 g+\tau_3(\theta\times g)}} \qquad (2)$$

where U_j denotes the observed response to item j, g is a dummy-coded group designator (in the current study $g = -.5$ for cancer patients and $g = .5$ for depressed patients), θ represents trait scores, τ_i ($i = 0, \ldots, 3$) denote logistic regression weights; and e is the base of the natural logarithm (*i.e.*, $e \approx 2.718$). As explained below, we are primarily interested in the magnitude and statistical significance of τ_2 and τ_3.

Admittedly, Eq. (2) is difficult to parse and the reader may find it difficult understanding how the various parameters identify item bias. Consequently, DIF researchers sometimes present this model in a statistically equivalent but somewhat clearer form. This latter model transforms the response variable into logits (log odds of item endorsement). In the logit metric, the log odds of an item endorsement can be expressed as a simple *linear* function:

$$\ln\left(\frac{P_j}{Q_j}\right) = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta \times g), \qquad (3)$$

where $Q_j = 1 - P_j$. From this perspective, it is easier to see that τ_2 captures the effect of group membership on the (transformed) item response probabilities after controlling for individual differences in trait level (θ) and for any possible group $\times$ trait interactions. In the present study, significant values of τ_2 would indicate that women with breast cancer and women with depression—with equivalent trait levels—had different probabilities of endorsing a BDI-II item. Note that if this was the case for many items that also showed clinically significant DIF then it would make little sense to compare these groups on BDI-II total scores (via a *t*-test or ANOVA, or to compare unconditional item endorsement rates) because observed scores in the two groups would correspond to different latent scores (psychometrically, the observed scores would be on different metrics). For instance, a score of 19 would not indicate the same level of depression for women with breast cancer and women with clinical depression. Summarizing these ideas more formally, significant τ_2 values, in combination with nonsignificant values of τ_3, provide evidence for uniform DIF.

Nonuniform DIF is operationalized as a significant interaction between trait scores and group membership ($\theta \times g$). Notice in Eq. (3) that this interaction is captured by τ_3. When τ_3 is large and significantly different from 0.0, the IRFs for the Reference and Focal groups cross, as they do in Fig. 1B.

Measuring DIF Effect Size: Or When DIF Makes a Difference

DIF analyses often require the computation of dozens of significance tests when they are conducted on moderate to large item pools. When conducting nonindependent significance tests, researchers may increase the likelihood of identifying false positive cases of DIF if the Type I error rate is not suitably controlled (Jodoin & Gierl, 2001). One way to keep the Type I error rate in check is to require DIF to be both statistically and clinically significant by supplementing the significance tests with a consideration of DIF effect size. We define clinical significance momentarily (note that we are not using the term clinically significant in the sense described by

Jacobson, Follette, & Revenstorf, 1984). First we describe how we tested statistical significance in our DIF study.

Several methods are available for testing statistical significance. In the present study we evaluated significance using a model comparison approach that controls Type I error rates by focusing on global DIF rather than uniform or nonuniform DIF [uniform and nonuniform DIF are easily assessed by considering the *t*-values of τ_2 and τ_3 in Eq. (2)]. To test for global DIF, we compared the relative fit of a model that allowed for DIF with that of a model specifying no DIF using a 2-degree of freedom chi-square test (see Rogers & Swaminathan, 1993 for details). By focusing on global DIF we performed half as many significance tests as that required in a study of uniform and nonuniform DIF.

We measured DIF effect size using a modified version of Raju's Noncompensatory DIF index (NCDIF; Raju, 1988; Raju, van der Linden, & Fleer, 1995). In this index, DIF is quantified by taking the squared, weighted difference between the IRFs in the Reference and Focal groups. Equation 4 expresses this idea algebraically. Notice in this equation that the squared differences between the response propensities ($[P_C(\theta_i) - P_D(\theta_i)]^2$) are weighted by the density of the trait levels in the group of cancer (C) patients.

$$\text{NCDIF}_j = \int_{-\infty}^{\infty} [P_{j,C}(\theta_i) - P_{j,D}(\theta_i)]^2 f_C\theta_i, d\theta, \quad (4)$$

where $P_{j,C}(\theta_i)$ denotes the probability of a keyed response on item j for cancer patients with trait level θ_i; $P_{j,D}(\theta_i)$ denotes the probability of a keyed response for depression patients with trait level θ_i, and $f_C\theta_i$ is the density of trait level θ_i for cancer patients.

In the present study we report $\text{NCDIF}^{1/2}$ (the square root of the NCDIF) rather than NCDIF so that our effect size measure is in a probability metric ([0,1]) rather than in a squared probability metric. An attractive feature of this index is that it focuses on probability differences within the trait range of the Focal group (i.e., cancer patients) rather than across the entire trait range. Recall that we are primarily interested in determining the effectiveness of the BDI-II as a measure of depressive symptoms in women with breast cancer.

Some readers may have questioned our use of a DIF model for binary items because BDI-II items are scored on a 4-point (ranging from 0 to 3) scale. Our decision was not arbitrary. Although DIF models are available for graded response items (Kim et al., 2002; Potenza & Dorans, 1995; Reise, Widaman, & Pugh, 1993; Santor et al., 1994a) we elected not to use these models in the present study because few cancer patients used the highest two response options on many BDI-II items. In this sample, the items were effectively binary, a situation that can arise in samples that are not highly populated with clinically depressed individuals.[5] In both samples, the BDI-II items were dichotomized by recoding all item responses greater than 1.00 to 1.00 (thereby measuring symptom presence versus absence).

METHOD

Sample

Participants were drawn from two samples, one comprised of women with newly diagnosed breast cancer and a second of women with a diagnosis of major depressive disorder. Both samples were recruited in randomized clinical trials to test the efficacy of psychological interventions; as such, they are representative of women seeking psychological help and willing to participate in a randomized study.

Clinically Depressed Sample

The depressed patients consisted of 294 women drawn from a recently completed placebo-controlled comparison between drugs (paroxetine) and cognitive or behavioral therapy. All patients were requesting treatment for depression and all measures were collected as part of the screening process prior to study entry. Although both men and women participated in the treatment study, only data from female patients are considered in this report. The first study was conducted jointly at Vanderbilt University and the University of Pennsylvania. Patients in this study had a mean age of 39.0 years ($SD = 11.75$), 33% were married or partnered, and they had a mean of 14.4 years of education ($SD = 2.29$). Nearly all

[5] Treating BDI-II items as binary can also be justified from a psychometric standpoint. Recent work by Santor and colleagues (Santor et al., 1994a, Santor, Ramsay, & Zuroff, 1994b; see Santor & Ramsay, 1998 for a review) suggests that some BDI item weights $(0, 1, \ldots 3)$ are not correctly ordered. In other words, for some items, the probability of endorsing higher response categories does not increase monotonically with depressive severity. Under this condition, scoring the items dichotomously actually increases our ability to understand the psychometric properties of the questionnaire (MacCallum, Zhang, Preacher, & Rucker, 2002) if we dichotomize at propitious thresholds.

of the patients were Caucasian (95.1%). These patients were supplemented with an additional group of women from a recently completed study of depression at the University of Washington. Inclusion criteria were identical to those used in the first study except that patients did not have to meet additional severity criteria beyond meeting criteria for DSM-IV major depression. The 159 women drawn from this second sample had a mean age of 39.19 years ($SD = 11.65$).

Breast Cancer Sample

The cancer patients consisted of 267 women diagnosed with Stage 0–III breast cancer from a randomized trial investigating the efficacy of two types of psychological group interventions (cognitive-behavioral and supportive expressive groups) designed to help women cope with the diagnosis and treatment of breast cancer. The interventions were designed to achieve several goals, including the reduction of symptoms of anxiety and depression. One intervention was based on cognitive-behavioral methods and included teaching relaxation, cognitive restructuring, and problem solving skills. The second intervention was based on a supportive expressive model of psychotherapy and involved supportive group sessions and journal writing. Importantly, data reported in this manuscript were collected prior to participation in these groups.

Fifteen percent of the women were diagnosed with Stage 0 cancer, 47.4% with Stage I, 32.5% with Stage II, and 4.9% with Stage III (women with Stage IV disease were not recruited for the study). These women had a mean age of 52.4 years ($SD = 10.6$), 74% were either married or partnered, and they had a mean of 14.6 years of education ($SD = 2.4$). Representative of the region in Northern New England from which the sample was recruited, 98.6% of the participants were Caucasian. Women were approached at their time of diagnosis and enrolled in the study. Depressive symptoms were measured on average 15 weeks after diagnosis as part of a larger packet of self-report measures. Patients were excluded if they had a history of organic or psychotic psychiatric disorder, if they had a diagnosis of metastatic (Stage IV) cancer, if they had a previous history of cancer, or if they had a significant comorbid medical condition (e.g., multiple sclerosis, diabetes). Patients were not included or excluded based on their current lifetime history of nonpsychotic Axis I DSM-IV disorders.

Based on structured psychiatric interviews, 1.9% of the breast cancer patients met criteria for current diagnosis of Major Depressive Disorder (MDD) and 20.3% met criteria for a lifetime history of MDD (Dausch et al., in press). In addition, 1.9% met criteria for a current diagnosis and 6.3% for a lifetime diagnosis of Dysthymic Disorder. The current and lifetime rates of MDD and Dysthymia in this sample are comparable to the rates reported on a representative community sample in the National Comorbidity Study (Kessler et al., 1994).

Data regarding adjuvant therapy for the sample were extracted from medical chart reviews. At the time that the data reported here were collected, 91 women (34%) had begun chemotherapy; those patients who were receiving chemotherapy at the time of completion of the BD-II had begun treatment on average 8.8 weeks prior ($SD = 5.8$). Eighty-five women (32%) had begun radiation therapy at the time of completion of the BDI-II; those receiving radiation therapy began treatment on average 6.9 weeks prior to completion of the BDI-II ($SD = 5.2$). Although 185 (73%) of women in this sample received hormonal therapy (tamoxifen), it was not possible to determine the start-date of this treatment from medical chart reviews for most women. Comparison of BDI-II total scores of those women who were ($M = 10.8$, $SD = 8.5$) and were not ($M = 9.0$, $SD = 7.2$) receiving radiation therapy indicated that these groups did not differ significantly. However, women who were receiving chemotherapy ($M = 12.2$, $SD = 8$) reported significantly more depressive symptoms than those who were not receiving chemotherapy ($M = 8.8$, $SD = 7.8$) $t(258) = 3.45$, $p < .001$. Cancer Stage (I–III) was significantly correlated with total BDI-II scores ($r = .21$, $p < .01$). When the presence or absence of chemotherapy was controlled, the partial correlation between stage and BDI-II total score was no longer significant ($r = .09$, $p = .11$). These findings are similar to those found in previous studies of women with newly diagnosed breast cancer (e.g., Compas et al., 1999).

Measures

The Beck Depression-II Inventory (BDI-II)

The BDI-II (Beck et al., 1996) was used in both studies to measure emotional, cognitive, and somatic symptoms of depression. Internal consistencies for the total score on the BDI-II were good for both

samples, with alphas of .86 for the sample of depressed women, and .92 for the sample of women with breast cancer.

Treatment of Missing Values

Few protocols in either sample had missing values on the BDI-II. Protocols were included in the final sample if they contained two or fewer missing items. For the few cases where it was necessary, missing data were imputed with an algorithm that was developed to treat missing data in DNA microarrays (Troyanskaya et al., 2001). One advantage of this algorithm over alternative density-based methods (Schafer & Graham, 2002) or ad hoc treatments (see Roth, 1994, for a review) is that it can be used in data sets that combine multiple populations. This feature is particularly important in the present study because we do not *assume* that cancer patients and depressed patients are drawn from a common population with respect to their performance on the BDI-II. On the contrary, this is a conjecture that we are interested in testing. The missing value algorithm was implemented with the EMV package (written by Raphael Gottardo) for the **R** programming environment (Ihaka & Gentleman, 1996).

RESULTS

Our analyses were conducted with two primary questions in mind. First, we wondered whether the BDI-II was an effective measure of depressive symptoms in women with breast cancer. To address this question we performed differential item functioning analyses (Holland & Wainer, 1993) on the BDI-II using samples of women with breast cancer (as a Focal group) and clinically depressed women (as a Reference group). Our second question considered the latent structure of the BDI-II in the breast cancer sample. Specifically, we wondered whether a psychometrically appropriate factor analysis of the BDI-II would shed light on the DIF analyses. Two important features of our analyses that differ from previous factor analyses of the BDI-II (Beck et al., 1996; Steer, Ball, Ranieri, & Beck, 1999; Steer, Kumar, Ranieri, & Beck, 1998) are that we used a procedure that is well suited to the analysis of clinical data (Waller, 2003) and we used a rotation method (Schmid & Leiman, 1957) that allowed us to simultaneously view general and group factors in a single solution. In the following sections we describe the DIF analyses and interpret the DIF findings in light of our factor analytic results.

Our first step in the DIF analyses was to develop an anchor test (Camilli & Shepard, 1994; Donoghue, Holland, & Thayer, 1993). In the present context, an anchor test is a relatively unbiased measure of the prominent depression factor that is measured by the BDI-II [θ in Eq. (2); A Monte Carlo study by Donoghue et al., 1993, found that including a small number of biased items in an anchor test decreases DIF effect sizes without vitiating the DIF analyses. More work is needed to study the effects of including a larger number of biased items.]. This measure was constructed by performing initial DIF analyses on the 21 BDI-II items and then purging the inventory of highly biased items (see Meredith & Millsap, 1992, for a discussion of why it is important to use relatively long anchor tests in observed score models). At the initial stage we used the unpurged total scores as a "first-pass" conditioning variable.

The preliminary DIF analyses identified six items as potentially biased (Items 1, 2, 3, 5, 13, & 17). Based on these preliminary results, the six offenders were removed from the BDI-II to create the anchor test. Note that when creating this test we summed the original, 4-point item responses rather than the binary item responses to maintain maximum variance.

In the next step, using our newly created matching variable (i.e., the anchor), we reran the DIF analyses by fitting the models and functions in Eqs. (3) and (4) to the item response data from the two samples. Table I summarizes the results and reports the logistic regression weights, the chi-square test of global (i.e., uniform and/or nonuniform) DIF, the $\text{NCDIF}^{1/2}$ measure of effect size, and a flag (*) marking items with clinically significant DIF. According to our criteria, clinically significant DIF was deemed present when: (i) the DIF χ^2 was significant at the .05 α level and (ii) the DIF effect size was greater than .10. Note that by jointly considering statistical significance and effect size we were also protected against high Type I error rates in the multiple tests (see Jodoin & Gierl, 2001 for a justification of this claim).

Our DIF analyses of the BDI-II revealed numerous items that functioned differently in the two samples. Notice that more than half of the items showed clinically significant DIF (71%) as previously defined. The mean DIF across the inventory was .25 (median = .21, SD = .21) whereas the mean DIF for the flagged items was .30 (median = .35, SD = .15). These results can be grasped more clearly by inspecting the DIF plots in Fig. 2.

Table I. DIP Results for 21 BDI-II Items: $P(U > 0)$

Item	τ_0	τ_1	τ_2	τ_3	χ^2_2	p-value	NCDIF$^{1/2}$
1. Sadness	−2.168	0.191	4.139	−0.148	14.947	0.001	0.48
2. Pessimism	−1.635	0.153	2.882	−0.13	11.219	0.004	0.41
3. Past failure	−1.010	0.098	4.816	−0.11	47.799	<.001	0.77
4. Loss of pleasure	−5.021	0.353	1.110	−0.043	0.617	0.734	0.01
5. Guilty feelings	−1.894	0.116	2.825	−0.063	21.945	<.001	0.36
6. Punishment feelings	−4.532	0.158	0.848	−0.064	6.583	0.037	0.01
7. Self-dislike	−3.934	0.240	2.840	−0.063	15.357	<.001	0.09
8. Self–criticalness	−3.828	0.187	2.249	−0.035	18.302	<.001	0.06
9. Suicidal thoughts or wishes	−3.505	0.115	2.664	−0.073	14.418	0.001	0.10
10. Crying	−2.721	0.188	−0.216	−0.044	9.099	0.011	0.02
11. Agitation	−2.155	0.152	2.126	−0.161	31.978	<.001	0.21
12. Loss of interest	−3.825	0.271	3.028	−0.110	6.766	0.034	0.10
13. indecisiveness	−1.958	0.155	3.467	−0.151	17.041	<.001	0.43
14. Worthlessness	−3.817	0.189	3.193	−0.102	14.078	0.001	0.10
15. Loss of energy	−0.391	0.201	2.014	−0.167	7.003	0.030	0.43
16. Changes in sleeping patterns	−1.450	0.226	2.182	−0.248	33.494	<.001	0.33
17. Irritability	−1.570	0.119	2.577	−0.099	11.431	0.003	0.38
18. Changes in appetite	−1.650	0.154	1.537	−0.132	19.566	<.001	0.21
19. Concentration difficulty	−2.680	0.243	1.126	−0.104	5.683	0.058	0.08
20. Tiredness or fatigue	−0.698	0.231	2.833	−0.297	29.555	<.001	0.54
21. Loss of interest in sex	−1.798	0.134	1.289	−0.119	23.623	<.001	0.16

Note. τ_0, τ_1, τ_2, & τ_3 are logistic regression weights from Eq. (3); NCDIF$^{1/2}$ is the square root of Raju's noncompensatory DIF index.

Several features of Fig. 2 merit comment. First, notice that the scores along the *x*-axis range from 0 to 63 even though the DIF analyses were conducted on a test with a raw score range of 0 to 45 (the anchor test included 15 items, each of which had a maximum score of 3). For ease of interpretation we rescaled the anchor test to the more familiar BDI-II metric using a regression linking procedure (to avoid metric bias, the regression equation was developed on the clinically depressed sample and then applied to both samples). Also notice that each subfigure includes "rug plots" on the bottom (*x*-axis) and top axes of the figure. The lower rug plot displays the score distribution for the breast cancer sample whereas the upper rug plot shows the score distribution for the clinically depressed patients. These rug plots demonstrate that the women with clinical depression had higher BDI-II scores than the cancer patients. Although this finding was expected, we are not focusing on group mean differences. Rather, we are interested in the performance of the BDI-II items after trait level has been held constant. In other words, we are interested in determining whether the items tap the same construct in the two groups.

Several features of the DIF analyses comparing the responses of depressed patients and breast cancer patients are striking. To gain a better understanding of these findings we interpret the results in light of an exploratory 5-dimensional factor structure identified in the present breast cancer sample. In the next section we describe the factor analytic methodology as a prelude to using the factor analysis results to guide our interpretations of the DIF findings.

The Factor Structure of the BDI-II in Breast Cancer Patients

The Beck Depression Inventory has been a popular focus of factor analytic studies of depression (see Dozois & Covin, in press, for a review). An informal PsychInfo literature search uncovered 75 references with the keywords "BDI and factor analysis." Although many of these studies have been informative, to our knowledge no previous study has applied a factor analytic methodology that is appropriate for the ordered-categorical responses of the BDI and BDI-II. In the next section we describe why traditional factor analyses of Pearson correlations may hide important features of the BDI-II factor structure.

Many years ago, John Carroll (1961) noted that the strength of a Pearson product moment correlation is constrained to the extent that the marginal distributions of two variables differ. In nontechnical terms, when the shape of two histograms differ (more formally, when the densities differ) two variables cannot achieve a perfect correlation even though both

variables may measure the same construct without error. Carroll was not the first to discuss this point (see Wherry & Gaylord, 1944), however his treatment of the issue was notable for its elegance and thoroughness. Importantly, Carroll noted that when these ideas are applied to questionnaire items, different densities are characterized by different item skewness values. These differences in distribution shape can bias Pearson product moments which in turn can bias factor analytic results by producing so-called "difficulty factors." Difficulty factors are factors that relate to item difficulty rather than to item substance (see McDonald & Ahlawat, 1974, for an updated interpretation of this issue). To avoid difficulty factors, researchers were advised to avoid the Pearson product moment correlation and to replace this statistic with tetrachoric or polychoric correlations.

To investigate these ideas in the present study, we computed BDI-II item skewness coefficients using data from the cancer patients. Our analyses revealed sizable differences among the 21 items. The coefficients are reported in the final column of Table II. Notice in this table that Suicidal Thoughts (Item 9) has a skewness of +2.88 whereas Tiredness or Fatigue (Item 20) has skewness of −1.58. Punishment Feelings (Item, 6) has a skewness of +2.15 whereas Loss of Energy (Item 15) has a skewness of −1.55. These differences are of sufficient magnitude that they could bias the factor structure of the BDI-II. To avoid this potential bias, we performed our factor analyses using tetrachoric correlations rather than Pearson product moments (we used tetrachorics rather than polychoric correlations because the former are more stable in samples of the size at our disposal; items were dichotomized at their medians). We used MicroFACT 2.1 (Waller, 2003) to perform this analysis. The Scree plot from our initial analysis suggested that four or five factors may be needed to account for the tetrachoric correlations. Using this information as a guide, we extracted (from the tetrachoric correlation matrix), rotated, and interpreted factor solutions that ranged from 2 to 5 factors.

Importantly, we did not use a canned oblique (e.g., Promax) or orthogonal (e.g., Varimax) rotation algorithm in these analyses. Had we done so we would have broken up the hypothesized general factor that runs through all 21 items. We desired to keep this general factor intact while at the same time rotating the solution to a position that elucidated the additional group factors that were suggested by the Scree plot (see McDonald, 1985, p. 106–107 for a discussion of why common rotation algorithms should not be used in solutions that contain both general and group factors). Our solution to this problem was to use a Schmid and Leiman rotation (1957). The Schmid Leiman rotation is particularly well-suited for hierarchical factor models in which all items are presumed to load on a general factor and one or more group factors. An attractive feature of this rotation is that all factors are orthogonal. Thus, the item variances are cleanly partitioned into orthogonal components of general factor variance and group factor variance. In other words, the groups factors are uncorrelated among themselves and uncorrelated with the general factor.

In our opinion, the 5-factor solution provided the most interpretable structure for the BDI-II items. [We recognize that GV5 is difficult to interpret and may have resulted from sample specific variance. Nevertheless, because it is generally preferable to overextract rather than underextract factors (Cattell, 1978), we report the 5-factor solution for completeness.] This solution is reported in Table II. Notice in Table II that a higher dimensional solution provides a finer picture of the item structure than that afforded by the 2-dimensional solution reported in earlier studies. Although all items load strongly on the general dimension in our solution, the group factors contribute significant variance. In other words, the group factors—in conjunction with the aforementioned DIF results—indicate that dimensions in addition to general depression, influence item response behavior on the BDI-II in women with breast cancer. For instance, *Gr2* appears to tap individual differences in vegetative symptoms (above that influenced by general depression); *Gr3* captures the Negative Self-Evaluation factor that has been identified in factor analyses of lexically derived trait descriptors (Waller, 1999); *Gr4* taps concentration difficulties, and *Gr5*—which may be the least stable factor—contrasts agitation with suicidal thoughts.

A general finding from the factor analytic findings in Table II is that DIF is not confined to one or more of the group factors. On the contrary, items with substantial DIF are found on all of the groups factors in columns 2–5 of the table. Moreover, several group factors have no items with clinically insignificant DIF. This latter finding is particularly disappointing because it suggests that it would be difficult to bowdlerize the BDI-II and create an unbiased instrument in samples with breast cancer patients.

Interestingly, items on *Gr3* share a common theme of negative cognitions about the self—feelings of past failure (Item 3), guilt (Item 5), self-dislike (Item 7), self-criticalness (Item 8), and worthlessness

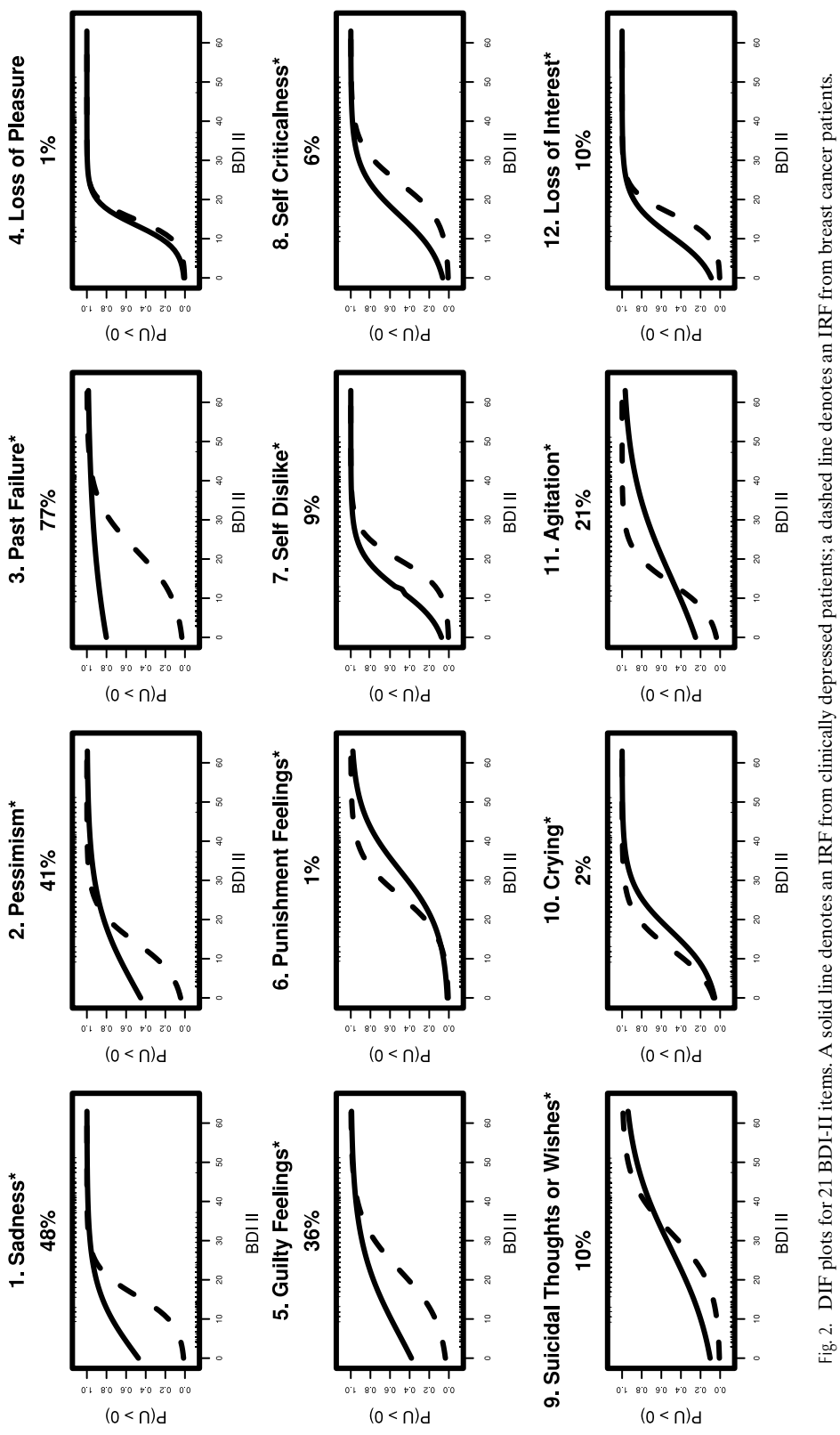

Fig. 2. DIF plots for 21 BDI-II items. A solid line denotes an IRF from clinically depressed patients; a dashed line denotes an IRF from breast cancer patients.

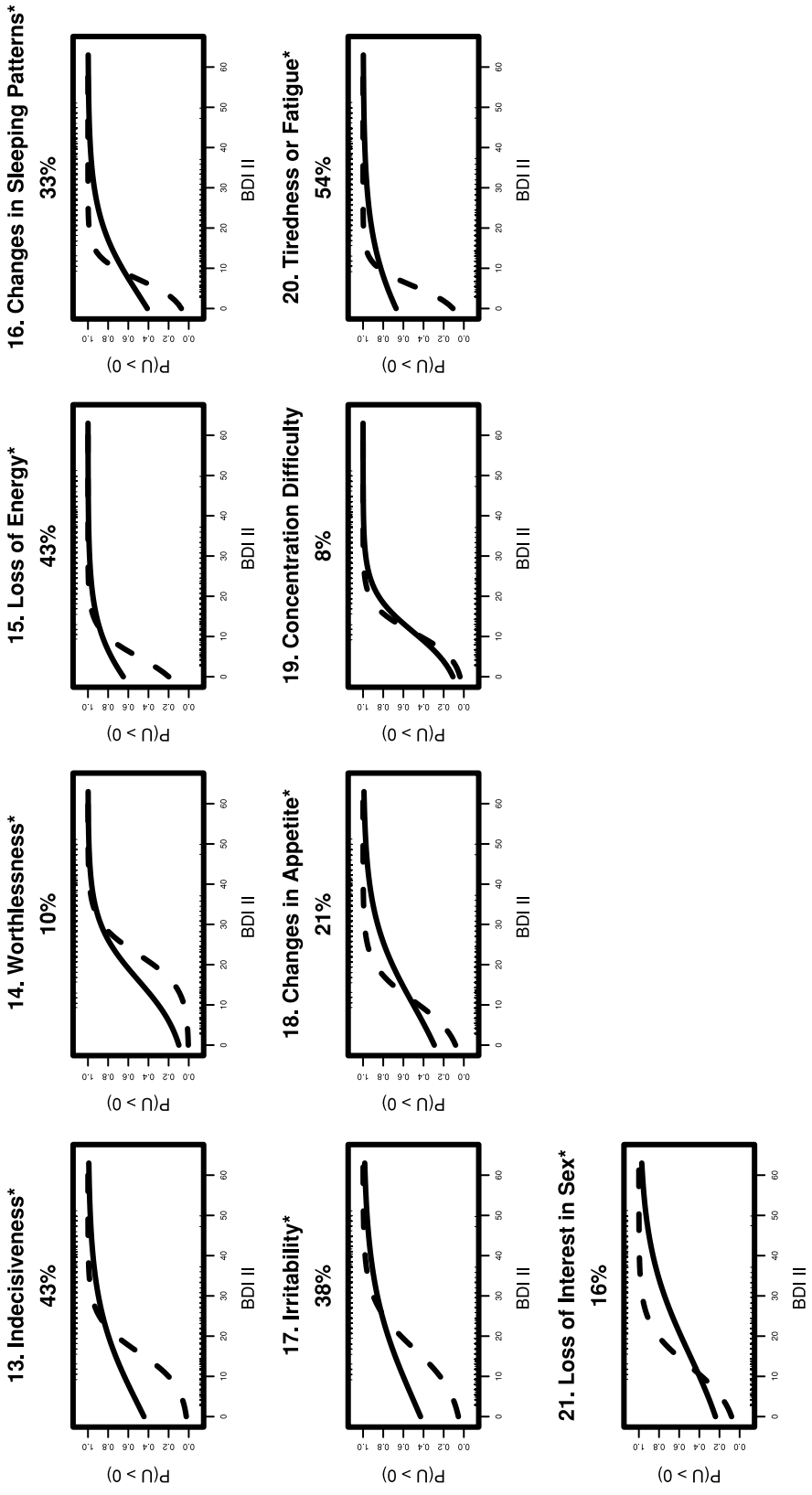

Fig. 2. Continued.

Table II. BDI-II General and Group Factor Structure in Breast Cancer Patients

Items	*G*	*Gr1*	*Gr2*	*Gr3*	*Gr4*	*Gr5*	NCDIF$^{1/2}$	Skew[a]
9. Suicidal thoughts or wishes	**0.54**	**−0.44**	0.10	−0.03	−0.07	0.32	.10	2.88
2. Pessimism	**0.75**	**−0.43**	0.16	0.08	−0.12	−0.07	.41	0.27
1. Sadness	**0.76**	**−0.42**	−0.02	0.00	0.09	−0.07	.48	0.62
20. Tiredness or fatigue	**0.56**	0.06	**−0.69**	0.12	−0.01	−0.25	.54	−1.58
15. Loss of energy	**0.44**	0.08	**−0.64**	0.15	−0.16	−0.07	.43	−1.55
18. Changes in appetite	**0.45**	0.07	**−0.59**	−0.27	0.00	0.19	.21	−0.22
16. Changes in sleeping patterns	**0.56**	−0.11	**−0.55**	0.01	0.07	0.00	.33	−1.04
8. Self-criticalness	**0.64**	0.11	−0.02	**−0.66**	−0.14	−0.05	.06	1.89
5. Guilty feelings	**0.57**	0.02	0.01	**−0.63**	0.06	−0.07	.36	1.23
3. Past failure	**0.60**	−0.14	0.13	**−0.51**	0.01	0.04	.77	1.25
14. Worthlessness	**0.73**	−0.04	0.13	**−0.51**	−0.15	−0.14	.10	1.75
13. Indecisiveness	**0.74**	−0.03	−0.03	−0.19	**−0.60**	0.11	.43	0.57
19. Concentration difficulty	**0.73**	−0.09	−0.21	0.05	**−0.40**	−0.05	.08	−0.22
17. Irritability	**0.66**	0.03	−0.17	−0.13	0.09	**−0.60**	.38	0.59
4. Loss of pleasure	**0.81**	−0.26	−0.13	−0.04	−0.13	−0.07	.01	0.52
6. Punishment feelings	**0.59**	−0.20	0.01	**−0.30**	0.11	−0.06	.01	2.15
11. Agitation	**0.70**	0.00	−0.13	−0.21	−0.04	**−0.39**	.21	0.24
7. Self-dislike	**0.76**	−0.18	0.07	**−0.28**	−0.07	−0.15	.09	1.30
10. Crying	**0.59**	−0.21	−0.10	0.04	−0.01	−0.17	.02	0.01
12. Loss of interest	**0.76**	−0.21	−0.10	−0.07	−0.27	0.03	.10	0.77
21. Loss of interest in sex	**0.50**	−0.25	−0.24	0.04	0.17	−0.07	.16	0.02

Note. *G* = general factor; *Gr*1–*Gr*5 = group level factors. Loadings ≥ |.30| in boldface.
[a]Item skewness computed in sample of cancer patients.

(Item 14). In previous factor analyses of lexically-derived trait descriptors, Tellegen and Waller (1987; Waller, 1999; Waller & Zavala, 1993) labelled this factor Negative Evaluation and they suggested that it is poorly measured in so-called Big Five factor structures. These are the items that might be expected to reflect more stable cognitive propensities that predispose psychiatric patients to depression (Beck, 1991). In a cognitive model of depression, these cognitive propensities are said to interact with negative life events to produce the onset of depression. Psychiatric patients with a history of depression are likely to be remain elevated on these items even when not currently depressed, whereas persons who lack such stable predisposition are likely to show elevations on these items only under states of extreme stress; e.g., after learning that they have cancer (Hollon, 1992). The one item in this cluster that showed the most prominent differential elevation among cancer patients at higher levels of depression was the sense of being punished (Item 6). On most of the other items in this cluster differences between the two samples were largely a function of cancer patients being less likely to endorse the cognitive items at lower levels of depression.

The two samples also differed on suicidal thoughts (Item 9), another item reflecting negative feelings or thoughts about the self. DIF analyses show that breast cancer patients were less likely to endorse suicidal thoughts than psychiatric patients at lower levels of depression, but slightly more likely to do so as depression levels rose. The same basic pattern held for irritability (Item 17) and two items reflecting aspects of cognitive functioning (indecisiveness and concentration difficulties). It is not clear why the latter should be the case. On the one hand, learning that one had cancer might be expected to be enough of a shock to interfere with ongoing decision-making and concentration (as suggested at higher levels of depression). However, if that were so, one would expect to see similar elevations in item endorsement at lower levels of depression, which was not the case.

DISCUSSION

Results of these initial analyses using DIF to compare the responses on the BDI-II of breast cancer patients with those of clinically depressed patients are provocative and suggest this method has considerable promise for increasing understanding of the responses of medically ill patients on this instrument. Patterns of items in which there was significant and nonsignificant DIF were identified in these samples. At the most general level, 15 of 21 BDI-II items showed significant DIF in the comparison of the breast cancer and depressed samples, suggesting that responses of

these samples to the majority of items do not reflect the same underlying construct.

There were pronounced and consistent differences between the two samples in the analyses of items representing negative cognitions about the self (e.g., worthlessness, self-dislike, punishment feelings). Compared with depressed patients, breast cancer patients were less likely to endorse negative beliefs about the self at low levels of depressive symptoms suggesting that negative cognitions about the self appear to be related to different factors in breast cancer patients than in depressed patients (Beck, 1991). Negative cognitions about the self appear to be more stable in the depressed patients and independent of depressive symptoms (Hollon, 1992). In contrast, in the cancer patients, negative cognitions about the self are more pronounced in the presence of higher levels of depression. Thus, these negative cognitions in cancer patients could be activated by the presence of an external stressor, in this case the diagnosis and treatment. In the depressed patients, these cognitions may be present irrespective of external sources of stress.

Closer examination of the patterns of DIF is instructive and will be useful in generating hypotheses for future research on the nature of depressive symptoms in cancer patients. For example, Item 3 (feelings of past failure) and Item 11 (agitation) both showed significant DIF but the patterns for the breast cancer and depressed samples were quite different on these two items. Depressed patients were highly likely to report feeling as if they failed in the past regardless of their total level of depressive symptoms. In contrast, the likelihood that cancer patients felt that they have been a failure was relatively low at low levels of total depressive symptoms but increased significantly at moderate to high levels of symptoms. This is consistent with previous studies that have shown that negative cognitions involving self-blame are linearly associated with higher symptoms of depression (and anxiety) in cancer patients (Glinder & Compas, 1999; Malcarne, Compas, Epping-Jordan, & Howell, 1995). In contrast, feelings of agitation and restlessness were less likely for breast cancer patients than for depressed patients at low levels of depressive symptoms. These patterns (and those observed on other items as well) highlight the need to investigate what factors other than total depressive symptoms may be leading to endorsement of specific symptoms in breast cancer patients.

In summary, we used modern psychometric methods for differential item functioning (Holland & Wainer, 1993; Raju et al., 1995) and item level factor analysis (Waller, 2003) to investigate the latent structure of the BDI-II in samples of women with breast cancer and women with clinical depression. Our analyses revealed many differences at both the item and factor scale levels; thus they have important implications for depression assessment in women with breast cancer. Most notably, they suggest that researchers and practitioners should apply caution when interpreting the BDI-II in breast cancer patients. We are currently collecting data from individuals with other forms of cancer to investigate the generalizability of these results.

ACKNOWLEDGMENTS

The authors wish to thank Drs. Judy Garber, Andrew Tomarken, David Cole, and other members of the Vanderbilt Depression Study Group for helpful comments on earlier drafts of this manuscript.

REFERENCES

American Cancer Society. (2003). *Cancer Facts and Figures 2003*. Atlanta, GA: American Cancer Society.

Beck, A. T. (1991). Cognitive therapy: A 30-year retrospective. *American Psychologist*, *46,* 368–375.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory: Manual* (2nd ed.). New York: The Psychological Corporation.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12,* 261–280.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased items*. Thousand Oaks, CA: Sage.

Carroll, J. B. (1961). The Nature of the data, or how to choose a correlation coefficient. *Psychometrika*, *26,* 347–372.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practices*, *17,* 31–44.

Compas, B. E., & Leucken, L. (2002). Psychological adjustment to breast cancer. *Current Directions in Psychological Science*, *11,* 111–114.

Compas, B. E., Stoll, M. F., Thomsen, A. H., Oppedisano, G., Epping-Jordan, J. E., & Krag, D. N. (1999). Adjustment to breast cancer: Age-related differences in coping and emotional distress. *Breast Cancer Research and Treatment*, *1233,* 1–9.

Croyle, R. T., & Rowland, J. H. (2003). Mood disorders and cancer: A National Cancer Institute Perspective. *Biological Psychiatry*, *54,* 192–194.

Dausch, B., Compas, B. E., Beckford, E., Luecken, L., Anderson-Hanley, C., Sherman, M., et al. (2004). Rates and correlates of DSM-IV diagnoses in women newly diagnosed with breast cancer. *Journal of Clinical Psychology in Medical Settings*, *11*(3), 159–169.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential Item Function* (pp. 137–166). Hillsdale, NJ: Erlbaum.

Dozois, D. J., & Covin, R. (in press). The Beck Depression Inventory-II, Beck Hopelessness Scale, and Beck Scale for Suicide Ideation. In M. Hersen (Ser. Ed.), D. L. Segal, & M. Hilsenroth (Vol. Eds.), *Comprehensive handbook of psychological assessment: Vol. 2. Personality assessment and psychopathology*. New York: Wiley.

Frasure-Smith, N., Lesperance, F., Juneau, M., Tlarijic, M., & Bourassa, M. G. (1999). Gender, depression, and one-year prognosis after myocardial infarction. *Psychosomatic Medicine*, *61,* 26–37.

Glinder, J., & Compas, B. E. (1999). Self-blame and psychological adjustment to breast cancer. *Health Psychology*, *18,* 1–9.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hollon, S. D. (1992). Cognitive models of depression from a psychobiological perspective. *Psychological Inquiry*, *3,* 250–243.

Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, *5,* 299–314.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15,* 336–352.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14,* 329–349.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Esheman, S., et al. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. *Archives of General Psychiatry*, *51,* 8–19.

Kim, Y., Pilkonis, P. A., Frank, E., Thase, M. E., & Reynolds, C. F. (2002). Differential functioning of the Beck Depression Inventory in late-life patients: Use of item response theory. *Psychology and Aging*, *17,* 379–391.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26,* 457–477.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7,* 19–40.

Malcarne, V. L., Compas, B. E., Epping-Jordan, J. E., & Howell, D. C. (1995). Cognitive factors in adjustment to cancer: Attributions of self-blame and perceptions of control. *Journal of Behavioral Medicine*, *18,* 401–417.

McDaniel, J. S., Musselman, D. L., Porter, M. R., Reed, D. A., & Nemeroff, C. B. (1995). Depression in patients with cancer: Diagnosis, biology, and treatment. *Archives of General Psychiatry*, *52,* 89–99.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, New Jersey: Erlbaum.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, *27,* 82–99.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57,* 289–311.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17,* 297–334.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, *19,* 23–37.

Raison, C. L., & Miller, A. H. (2003). Depression in cancer: New developments regarding diagnosis and treatment. *Biological Psychiatry*, *54,* 283–294.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53,* 495–502.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures for differential functioning of items and tests. *Applied Psychological Measurement*, *19,* 353–368.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114,* 552–566.

Ritterband, L. M., & Spielberger, C. D. (2001). Depression in a cancer patient population. *Journal of Clinical Psychology in Medical Settings*, *8,* 85–93.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17,* 105–116.

Rosenberg, S. D., Goodman, L. A., Osher, F. C., Swartz, M. S., Essock, S. M., Butterfield, M. L., et al. (2001). Prevalence of HIV, hepatitis B, and hepatitis C in people with severe mental illness. *American Journal of Public Health*, *91,* 31–37.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, *47,* 537–560.

Santor, D. A., & Coyne, J. C. (2001). Examining symptom expression as a function of symptom severity: Item performance on the Hamilton Rating Scale for Depression. *Psychological Assessment*, *13,* 127–139.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment*, *10,* 345–359.

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994a). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, *6,* 255–270.

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994b). Nonparametric item analysis of the Beck Depression Inventory: Examining item bias and response option weights in clinical and nonclinical samples, *Psychological Assessment*, *6,* 255–270.

Santor, D. A., Zuroff, D. C., Cervantes, P., Palacios, J., & Ramsay, J. O. (1995). Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychological Assessment*, *7,* 131–139.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7,* 147–177.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22,* 53–61.

Shealy, R., & Stout, W. (1993). An item response theory model for test bias. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale, NJ: Erlbaum.

Spijker, A. V., Trijsburg, R. W., & Duivenvoorder, H. J. (1997). Psychological sequelae of cancer diagnosis: A meta-analytic review of 58 studies after 1980. *Psychosomatic Medicine*, *59,* 280–293.

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.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27,* 361–370.

Tellegen, A., & Waller, N. G. (1987). *Reexamining basic dimensions of natural language trait descriptors*. Paper presented at the 95th annual meeting of the American Psychological Association, New York, New York.

Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago: Scientific Software International.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group differences: The concept of item bias. *Psychological Bulletin*, *99,* 118–128.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17,* 520–525.

Waller, N. G. (1998a). LINKDIF: An S-Plus routine for linking item parameters and calculating IRT measures of differential functioning of items and tests. *Applied Psychological Measurement*, *22,* 392.

Waller, N. G. (1998b). EZDIF: A program for the analysis of uniform and nonuniform differential item functioning. *Applied Psychological Measurement*, *22,* 391.

Waller, N. G. (1999). Evaluating the structure of personality. In C. Robert Cloninger (Ed.) *Personality and psychopathology* (pp. 155–200). Washington, DC: American Psychiatric Press.

Waller, N. G. (2003). *WinMicroFACT 2.1: A microcomputer factor analysis program for ordered polytomous data and mainframe sized problems*. St. Paul, MN: Assessment Systems Corporation.

Waller, N. G., Thompson, J., & Wenk, E. (2000). Black–White differences on the MMPI: Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales. *Psychological Methods*, *5,* f25–f46.

Waller, N. G., & Zavala, J. (1993). Evaluating the Big Five. *Psychological Inquiry*, *4,* 131–134.

Wherry, R. J., & Gaylord, R. H. (1944). Factor Pattern of test items and tests as a function of the correlation coefficient: Content, difficulty, and constant error factors. *Psychometrika*, *9,* 237–244.