Psychological Assessment
2011, Vol. 23, No. 1, 193–202

1040-3590/11/$12.00 DOI: 10.1037/a0021292

An Evaluation of the Brief Symptom Inventory–18 Using Item Response Theory: Which Items Are Most Strongly Related to Psychological Distress?

Rob R. Meijer and Rivka M. de Vries
University of Groningen

Vincent van Bruggen
Dimence, Almelo, the Netherlands

The psychometric structure of the Brief Symptom Inventory–18 (BSI-18; Derogatis, 2001) was investigated using Mokken scaling and parametric item response theory. Data of 487 outpatients, 266 students, and 207 prisoners were analyzed. Results of the Mokken analysis indicated that the BSI-18 formed a strong Mokken scale for outpatients and prisoners, indicating strong unidimensionality. For students, only the depression and anxiety items formed a medium Mokken scale. Parametric item response theory analyses showed that the best discriminating items came from the depression and anxiety subscales.

Keywords: Brief Symptom Inventory, Mokken scaling, item response theory, psychological distress

The Brief Symptom Inventory–18 (BSI-18; Derogatis, 2001) is a widely used self-report questionnaire that measures general psychological distress. It is the briefest and latest version in a series of instruments designed by Derogatis (Derogatis, 1983; Derogatis & Melisaratos, 1983; Derogatis, 2001) to measure general distress. The questionnaire consists of 18 descriptions of physical and emotional complaints; respondents are asked to indicate on a scale from 0 (*not at all*) through 4 (*very much*) to what extent they are troubled by the complaints. Table 1 shows the item content.

The BSI-18 is a shortened version of the BSI, which consists of 53 items distributed over nine subscales: Somatization, Obsessive-Compulsive Disorder, Interpersonal Sensitivity, Depression, Anxiety, Hostility, Phobic Anxiety, Paranoid Ideation, and Psychoticism. Piersma, Boes, and Reaume (1994), among others, found that responses to all items can be described by a unidimensional model. The BSI was reduced to the BSI-18 to decrease the average completion time and to improve its structural validity (Derogatis, 2001). According to Derogatis (2001), the structural validity has improved because the reduced scale is composed of only three dimensions—namely, somatization, depression, and anxiety—which together are more homogeneous than other dimensions from previous instruments, both conceptually and empirically. Each subscale of the BSI-18 contains six items from the three corresponding subscales of the BSI. A total score over all items can be calculated representing general distress, which is highly correlated with the total score from the BSI ($r > .90$; Andreu et al., 2008; Durá et al., 2006; Galdón et al., 2008).

Although some authors (e.g., Piersma et al., 1994) have claimed that one dimension underlies the item scores of the BSI, Derogatis (2001) and several researchers claim that the BSI-18 has a multidimensional structure. Piersma et al. (1994) administered the complete BSI to 217 adults and 188 adolescents at admission and discharge from a private psychiatric hospital. Principal components factor analysis revealed that most variance between dimension scores was accounted for by one unrotated factor. In contrast, Galdón et al. (2008), using a sample of 175 breast cancer patients, found three dimensions underlying the BSI-18. These dimensions corresponded to the hypothesized subscales somatization, depression, and anxiety. The same structure was found by Durá et al. (2006) and Recklitis et al. (2006). Durá et al. (2006) investigated 114 patients with temporomandibular disorders. Recklitis et al. (2006) investigated a sample of 14,193 adult survivors of childhood cancer. The mean scores on the items varied between .12 (Item 17, "suicidal thoughts") and .77 (Item 6, "feeling tense"). Thus, the general distress level was low in this sample. Andreu et al. (2008) also reported multidimensionality but preferred a four-dimensional structure, where the anxiety dimension was split into a general anxiety dimension and a panic dimension. The sample in the Andreu et al. (2008) study consisted of 200 outpatients with psychological symptomatology. Fifty-two percent of the outpatients were recruited from a private psychology clinic, whereas 48% came from public services. Fifty-five percent were women and the remaining 45% were men. The majority of the sample (53.4%) were diagnosed with anxiety disorders; 32% were diagnosed with major depression disorders. The mean item scores on the item varied between .94 (Item 1, "dizziness") and 2.51 (Item 3, "feeling blue") and were considerably higher than in the Recklitis et al. (2006) study. According to Andreu et al. (2008), the four-dimensional structure may be specific for patients with psychiatric disorders, in contrast to patients with medical problems for whom the threat caused by the medical condition might homogenize their experience of fear.

In the studies cited above, a multidimensional structure was preferred to a unidimensional structure based on model fit. However, when we consider the fit indices reported in these studies, differences between unidimensional and multidimensional models

This article was published Online First January 31, 2011.

Rob R. Meijer and Rivka M. de Vries, Department of Psychometrics and Statistics, University of Groningen, Groningen, the Netherlands; Vincent van Bruggen, Dimence, Almelo, the Netherlands.

Correspondence concerning this article should be addressed to Rob R. Meijer, University of Groningen, Department of Psychometrics and Statistics, Grote Kruisstraat 2/1, 9712 TS Groningen, the Netherlands. E-mail: r.r.meijer@rug.nl

Table 1
Item Descriptives, Item-Total Correlation With Subscale (r_s), Item-Total Correlation With Total Scale (r_t), Rho, Scale H, Item H Value Within Subscale (H_{is}), and Item H Value Within Total Scale (H_{it}) in the Sample With Clinical Patients

Item	Item number	*M*	*SD*	r_s	r_t	H_{is}	H_{it}
Somatization (ρ = .86, *H* = .53)							
Feeling weak	16	1.19	1.36	.79	.78	.55	.57
Nausea	7	1.01	1.35	.77	.70	.53	.50
Numbness	13	0.88	1.22	.76	.66	.54	.50
Faintness	1	0.83	1.06	.73	.68	.52	.49
Trouble getting breath	10	0.69	1.14	.75	.67	.51	.48
Pains in chest	4	0.78	1.11	.78	.63	.53	.46
Depression (ρ = .90, *H* = .64)							
Feeling blue	8	1.78	1.29	.91	.85	.75	.65
Feeling no interest in things	2	1.33	1.34	.85	.82	.62	.58
Feeling lonely	5	1.60	1.34	.82	.68	.63	.46
Feeling hopeless about future	14	1.46	1.32	.87	.78	.67	.54
Feeling of worthlessness	11	1.21	1.30	.80	.71	.59	.52
Suicidal thoughts	17	0.49	0.86	.62	.54	.51	.43
Anxiety (ρ = .89, *H* = .63)							
Feeling tense	6	1.77	1.30	.86	.83	.67	.63
Nervousness	3	1.55	1.22	.81	.73	.65	.54
Feeling fearful	18	1.29	1.24	.83	.77	.64	.57
Spells of panic	12	1.15	1.27	.84	.79	.65	.56
Suddenly scared	9	1.30	1.30	.82	.74	.63	.56
Feeling restless	15	1.07	1.22	.68	.63	.50	.45

were not always compelling. For example, Recklitis et al. (2006) found a three-factor structure when conducting a confirmatory factor analysis. But when fitting a hierarchical model with depression, anxiety, and somatization as first-order factors and general distress as a second-order factor, they found very high correlations between the first-order factors and the second-order factors. Somatization correlated $r = .98$ with general distress; depression and anxiety correlated $r = .79$ and $r = .74$ with general distress, respectively. These high correlations suggest that there is a strong general dimension here. In sum, from the existing literature it is unclear whether the BSI-18 items form a unidimensional scale or whether different scales can be distinguished.

Thus far, the dimensionality of the BSI-18 has been investigated using models based on classical test theory (CTT) and factor analysis. In the present study we first investigated the dimensionality of the BSI-18 using Mokken scaling (Sijtsma & Molenaar, 2002), which is a nonparametric item response theory (IRT; Embretson & Reise, 2000) technique. Second, we used a parametric IRT model to obtain more detailed information about the quality of the items across different psychological distress levels.

IRT is a collection of statistical models that can be used to evaluate and construct psychological tests and questionnaires. Although there are similarities between IRT, CTT, and factor analysis, IRT has a number of advantages when evaluating the psychometric quality of a scale (see, e.g., Egberink & Meijer, 2010; Reise & Waller, 2009; Santor & Ramsay, 1998). An important advantage of IRT is that to judge the quality of an item, one can obtain the item information function, which shows how much psychometric information (a number that represents an item's ability to differentiate between people) the item provides at each trait level (such as psychological distress). Different items can provide different amounts of information in different ranges of a given latent trait. Item and scale information are analogous to CTT's item and test reliability. An important difference, however, is that under an IRT framework, information (precision) can vary depending on where an individual falls along the trait range, whereas in CTT, the scale reliability (precision) is assumed to be the same for all individuals, regardless of their raw-score levels. As some authors have discussed (Recklitis et al., 2006), it is important for studies to verify that the BSI-18 is sensitive to change in a population so that it can monitor clinical course or the outcome of an intervention. When measuring change, it is thus important to have knowledge about the measurement precision conditional on the latent trait.

Furthermore, we extend the literature by investigating the psychometric structure in a sample of outpatients with anxiety and unipolar and bipolar disorders, a sample of students, and a sample of prisoners. When a scale is applied to populations with different characteristics, the psychometric properties of the scale may vary. Most studies on the BSI-18 have been conducted with medical samples, in particular with samples of cancer patients (Durá et al., 2006; Galdón et al., 2008; Recklitis et al., 2006). An exception is Andreu et al. (2008), who analyzed data of a sample consisting of outpatients with psychiatric disorders.

Our aim is to discover (a) how BSI-18 items are functioning in different populations, (b) which items have the strongest relation to the constructs being measured, and (c) whether we can scale persons on the basis of the 18 items of the BSI-18.

Method

Samples

The first sample consisted of 487 outpatients with anxiety and depression disorders. 53% was diagnosed with anxiety disorder, 37% was diagnosed with unipolar or bipolar disorder, the remaining disorders were unknown. The sample was 62.5% female and 37.5% male. Mean age was 34.4 (*SD* = 10.9). Data were obtained

as part of a psychological assessment and treatment program in the east of the Netherlands. The items were used as a screening instrument for patients before starting treatment in a psychological clinic. Treatments that the patients received were mainly based on cognitive behavioral approaches. Note that this sample has some similarities with the sample analyzed by Andreu et al. (2008) and that it consists of persons for whom the BSI-18 is often used.

A second sample consisted of 266 psychology students (29% male, 71% female; mean age 22.2, $SD = 4.1$) that filled out the BSI-18 for screening purposes. These students were not selected on the basis of reporting psychological problems or on the basis of elevated distress levels. Instead, the BSI-18 was used to screen for students with potential psychological problems. Because we expect that in this population the general distress level is lower than the distress level in medical or clinical samples, this sample is useful to obtain information about the psychometric structure for populations with low general distress levels.

A third sample consisted of 207 prisoners (94% male, 6% female; mean age 34.10, $SD = 9.5$). Their self-reported ethnicity was 51% African descent, 25% White, 4% Hispanic, 3% Asian; for the remaining prisoners, ethnicity was unknown. Data were collected at different prisons in the Netherlands as part of forensic research. All testing was done by forensic psychologists at the various institutions, and all the prisoners were tested on intake.

Each sample was analyzed separately. We did not combine samples because that would lead to misleading results. Waller (2008) showed that reliability coefficients and related indices (such as the *H* values we use) are severely biased when samples are commingled—that is, when they are drawn from multiple populations. In most cases the estimates are inflated. Furthermore, results based on samples from two or more populations (e.g., combined community and clinical samples) will yield a mean on the latent trait that is difficult to interpret (cf. Reise & Waller, 2009).

Item Response Theory

For dichotomous items, unidimensional IRT is based on the assumption that a person's performance on a test item can be predicted by the interplay between a latent trait theta (θ) and the item characteristics, such as item discrimination and item difficulty (e.g., Hambleton, Swaminathan, & Rogers, 1991). The relationship between item performance and the trait level theta can be described by a monotonically increasing function, which is called the *item characteristic function*, the *item characteristic curve,* or the *item response function* (IRF). Let $P_i(\theta)$ be the probability of a positive response (i.e., a correct answer or the agreement with a specific statement) on item i for a given level of θ. Then the core assumption states that when the trait level theta increases, the probability of a positive item response $P_i(\theta)$ also increases. For polytomous items, this assumption is made at the level of *item steps*, which are the transitions from one answering category to the next. For example, subjects choosing Category 2 on a 4-point scale have a score of 1 on the first two item steps (from 0 to 1 and from 1 to 2) and a score of 0 on the second two item steps (from 2 to 3 and from 3 to 4). A distinction can be made between parametric and nonparametric IRT models.

Parametric IRT. Parametric IRT models describe the relationship between the probability of a positive response and the latent trait level (the IRF) by means of a parametric (e.g., logistic) function. An often-used model for dichotomous items is the two-parameter logistic model (Embretson & Reise, 2000). This model contains, in addition to the latent trait parameter theta, two parameters representing item characteristics. One parameter is the item location or item difficulty β_i, which is the location on the theta scale for which $P_i(\theta) = .5$. The other parameter is the discrimination parameter or α_i parameter. The α_i parameter is the steepness of the IRF at the item difficulty level. In practice, α_i ranges from 0 (flat IRF) to 3 (steep IRF). Items with a larger α_i parameter are more useful for separating examinees near a trait level.

For polytomous items, extensions of the models for dichotomous items have been developed. An extension of the two-parameter logistic model is the graded response model of Samejima (1969) for ordinal answering categories, which models the probability that an examinee responds in a particular answering category n. The model contains a discrimination parameter and a number of location parameters (denoted β_n) equal to the number of answering categories minus 1. As before, the discrimination parameter reflects the strength of the relationship between the item and the latent trait. The location parameter for a specific category β_n is the location on the theta scale for which the probability of scoring in this category or higher is .5. Together the location parameters reflect the spacing of the answering categories on theta.

In addition to the difficulty and discrimination parameters, a useful concept in describing the quality of the items is the *item information*. The item information is the inverse of the standard error of measurement, so more measurement error results in less information. In IRT the measurement error and information depend on theta, which is different from the single estimate of measurement error in CTT, which is assumed to be equal across different values of theta. The information an item provides about a person is higher when the item difficulty β_i is close to theta and when α_i is high. Once a valid model has been constructed, it can be used to estimate theta for specific persons on the basis of their test scores. Examples of applications of parametric IRT modeling can be found in, for example, Lambert et al. (2003), who investigated the Youth Self-Report; Teresi et al. (2000), who investigated the Comprehensive Assessment and Referral Evaluation diagnostic scale; and Emons, Meijer, and Denollet (2007), who investigated a questionnaire measuring Type D personality.

Nonparametric IRT. In contrast to parametric models, nonparametric models do not fully determine the IRFs (Hambleton, Swaminathan, & Rogers, 1991). Examples of nonparametric IRT models are the Mokken models (Sijtsma & Molenaar, 2002). The least restrictive Mokken model is the *monotone homogeneity model* (MMH model), which only requires that the relationship between $P_i(\theta)$ and θ is monotonely nondecreasing. That is, if for two persons a and b it holds that $\theta_a < \theta_b$, then it should also hold that $P_i(\theta_a) \leq P_i(\theta_b)$. A more restrictive Mokken model is the *double monotonicity model*, where the additional assumption of nonintersecting IRFs is made (e.g., Meijer, 2010).

Mokken models do not offer estimates of parameters like β_i and α_i, nor do they allow for point estimates of theta. However, several measures can be used to obtain an idea about the quality of the scale, such as the item proportion correct score reflecting item difficulty and scalability coefficients (H) reflecting discrimination power. Besides, at the scale level, H is also defined at the item(step)-pair level (H_{ij}) and item level (H_i) and can be expressed

in terms of observed versus expected number of Guttman errors or in terms of observed versus maximal possible covariance between items (for exact formulas, see, e.g., Sijtsma & Molenaar, 2002, pp. 51–58). For the interpretation of H, Sijtsma and Molenaar (2002, pp. 60) give the following guidelines. The scale H should be above .3 for the items to form a scale. When $.3 \leq H < .4$, the scale is considered weak; when $.4 \leq H < .5$, the scale is considered medium; and when $H \geq .5$, the scale is considered strong. In addition, although point estimates of theta are not possible, an estimated ordering of subjects by their theta values is possible using the number-correct score. Examples of applications of Mokken scaling in the typical performance domain can be found in, for example, Meijer and Baneke (2004), who showed the usefulness of Mokken scaling to analyze the MMPI depression scale; Moorer, Suurmeijer, Foets, and Molenaar (2001), who applied Mokken scaling to the Rand-36; and Meijer, Egberink, Emons, and Sijtsma (2008), who discussed the use of Mokken scaling to identify atypical response behavior.

Analysis

First, we used Mokken models to investigate the psychometric structure of the data. These models are excellent tools for a first exploration of the psychometric structure of test and questionnaire data. In contrast to parametric IRT models, Mokken models do not specify the exact relation between endorsing an item and the latent trait level; as a result, they are less restrictive to empirical data than parametric models. Mokken scale analyses were performed using the computer program Mokken Scale Analysis for Polytomous Items (MSP5.0; Molenaar & Sijtsma, 2000).

We started with running the TEST option in MSP5.0. This is a procedure where the researcher specifies which items form a scale. The three subscales as defined in the literature (somatization, depression, and anxiety) were analyzed separately, as well as together forming the total scale that measures general distress. The main focus was on the H values of the different scales and the total scale. Also, a reliability coefficient rho (ρ) was estimated for each scale, which is an unbiased estimate rather than a lower bound like Cronbach's alpha (Moorer et al., 2001).

In addition, the SEARCH option was applied, which is an exploratory procedure searching for unidimensional scales within a specified set of items. The procedure starts with the item pair consisting of the items i and j with the largest pairwise H value, H_{ij} (alternatively, the researcher can specify a start set of items). In the next step, one of the remaining items is selected that (a) correlates positively with each of the items already selected in the scale, (b) has an H_i with respect to the selected items significantly larger than 0 and also larger than a prespecified value c, and (c) maximizes the total H of the scale. This step is repeated until none of the items left meet the criteria for selection. Then the procedure starts again, now applied on the remaining items, if any, until no items are left. Some items may not reach the criteria for any scale and are left out. Thus c is a constant, and often $c = .3$ is used. The higher c is, the more confidence we have in the ordering of persons according to their total score.

The SEARCH procedure is useful for investigating the dimensionality of the scale. Sijtsma and Molenaar (2002, pp. 81–82) give the following guidelines for determination of the dimensionality. For unidimensional scales, the typical results with increasing c are (a) most or all items are on one scale, (b) one smaller scale is found, and (c) one or a few small scales are found and several items are excluded. In multidimensional scales, the typical results with increasing c are (a) most or all items are on one scale, (b) two or more scales are formed, and (c) two or more smaller scales are formed and several items are excluded. A strength of this procedure is that it removes most of the items that do not satisfy the MMH model; depending on the choice of the lower bound, it removes items that hardly contribute or contribute only modestly to the dimensional structure of the data. A drawback may be that this research algorithm is not a formal test of the MMH model. Sometimes an item may be rejected that shows a few local decreases in the IRF or has an increasing but relatively flat IRF.

Second, the graded response model was estimated using MULTILOG7 (Thissen, Chen, & Bock, 2003). The graded response model was estimated to obtain item and person parameters and to estimate the information curves for the subscales and total scale.

Results

Descriptive Statistics

Tables 1, 2, and 3 present the item means with their standard deviations, together with the item-total correlations for the subscales and the total scale (the item and scale H values are also presented in the table and will be discussed below) for the clinical, prisoner, and student samples, respectively. The items are clustered according to their theoretical assignment to the different dimensions in earlier research (e.g., Derogatis, 2001). Table 4 presents the intercorrelations between the subscales and the total scale scores for the three samples.

Clinical sample. In the clinical sample, the mean total score for the complete scale (which is also referred to as the Global Severity Index) equaled $M = 21.27$ ($SD = 16.02$). Compared with the mean item scores found in medical samples (cancer patients), which are generally below 1 (Durá et al., 2006; Galdón et al., 2008; Recklitis et al., 2006), the average item scores in this sample tended to be higher. That is, the subjects were more distressed. Except for one item ("suicidal thoughts"), all depression and anxiety items had means between 1.07 and 1.78. This is lower, however, than the average item scores found by Andreu et al. (2008) in a psychiatric sample, which ranged from 1.27 through 2.54 with two exceptions. Subscale reliabilities were high; all rho values were .86 or higher. The rho estimate of the total scale equaled .95. As shown in Table 1, both the item-total correlations for the subscales and the total scale were high. High correlations were also observed between the subscales and between the subscales and the total scale (Table 4, upper panel). Correlations between the subscales ranged from .63 to .76, and correlations between subscales and the total scale ranged from .88 to .92.

Prisoner sample. The mean total score for the complete scale was lower than in the clinical sample ($M = 16.34$, $SD = 14.20$). The rho estimate equaled .94. As shown in Table 2, the mean item scores on the somatization items were lower than on the depression and anxiety items. The item-total correlations for each subscale and for the total scale were high, although somewhat lower than for the clinical sample. Interesting was that the only item on which the prisoners scored higher than the patients

Table 2
Item Descriptives and Item and Scale Statistics in the Sample With Prisoners

Item	Item number	*M*	*SD*	r_s	r_t	H_{is}	H_{it}
Somatization (ρ = .84, *H* = .50)							
Feeling weak	16	0.72	1.19	.57	.64	.47	.52
Nausea	7	0.59	1.02	.61	.58	.49	.46
Numbness	13	0.68	1.09	.74	.67	.57	.51
Faintness	1	0.56	1.06	.59	.52	.49	.41
Trouble getting breath	10	0.24	0.63	.50	.47	.45	.42
Pains in chest	4	0.68	1.08	.62	.58	.50	.45
Depression (ρ = .88, *H* = .63)							
Feeling blue	8	1.73	1.59	.87	.86	.75	.59
Feeling no interest in things	2	0.88	1.27	.65	.58	.60	.49
Feeling lonely	5	1.78	1.51	.74	.63	.66	.51
Feeling hopeless about future	14	1.43	1.43	.76	.70	.66	.54
Feeling of worthlessness	11	0.89	1.19	.65	.64	.58	.48
Suicidal thoughts	17	0.50	0.85	.48	.47	.47	.44
Anxiety (ρ = .86, *H* = .61)							
Feeling tense	6	1.50	1.25	.72	.68	.66	.63
Nervousness	3	1.09	1.21	.73	.66	.64	.50
Feeling fearful	18	0.91	1.20	.67	.75	.58	.56
Spells of panic	12	0.54	1.06	.71	.66	.65	.53
Suddenly scared	9	0.72	1.08	.76	.76	.66	.58
Feeling restless	15	0.89	1.15	.56	.57	.50	.43

in the clinical sample was Item 5 ("Feeling lonely"). Intercorrelations between the scales were comparable with the clinical sample (Table 4, middle panel).

Student sample. The mean total score equaled M = 8.41 (*SD* = 7.83, ρ = .88) and was comparable with the mean score found in the Recklitis et al. (2006) study using cancer survivors (their mean score equaled 6.18 for the total sample). This may seem surprising, but Recklitis et al. (2006) already discussed that the low distress level in their study may be due to "improved coping skills or social supports . . . over time this develops into a form of resilience, making survivors less prone to psychological symptoms under subsequent stress" (p. 29). It is interesting that the item-total correlations for each subscale were, in general, somewhat lower than in the clinical and prisoner samples and that the item-total correlations for the total scale were considerably lower than in the clinical and prisoner samples, especially for the somatization items. Also, intercorrelations between the scales were lower than for the clinical and prisoner samples (Table 4, lower panel). In the samples there were few missing values (between zero and four per item). We used multiple imputation (Van Ginkel, Van der Ark, & Sijtsma, 2007) to obtain item scores for these items.

Table 3
Item Descriptives and Item and Scale Statistics in the Sample With Students

Item	Item number	*M*	*SD*	r_s	r_t	H_{is}	H_{it}
Somatization (ρ = .79, *H* = .32)							
Feeling weak	16	.74	.84	.50	.52	.36	.33
Nausea	7	.64	.91	.50	.44	.36	.29
Numbness	13	.23	.54	.40	.36	.31	.24
Faintness	1	.46	.73	.41	.44	.31	.29
Trouble getting breath	10	.21	.58	.39	.41	.31	.28
Pains in chest	4	.23	.54	.29	.28	.23	.19
Depression (ρ = .84, *H* = .54)							
Feeling blue	8	.56	.90	.77	.73	.63	.46
Feeling no interest in things	2	.39	.78	.62	.53	.53	.34
Feeling lonely	5	.74	.94	.63	.54	.55	.35
Feeling hopeless about future	14	.58	.87	.67	.53	.57	.34
Feeling of worthlessness	11	.37	.77	.58	.55	.50	.35
Suicidal thoughts	17	.13	.47	.37	.35	.38	.26
Anxiety (ρ = .80, *H* = .40)							
Feeling tense	6	.94	.97	.72	.64	.45	.42
Nervousness	3	.83	.97	.73	.54	.42	.35
Feeling fearful	18	.32	.66	.67	.51	.64	.33
Spells of panic	12	.22	.54	.71	.49	.41	.33
Suddenly scared no reason	9	.32	.66	.76	.47	.41	.31
Feeling restless	15	.48	.84	.56	.46	.34	.30

Table 4
Correlations Between the Subscales and the Total Score for Different Samples

Sample	Somatization	Depression	Anxiety
Clinical			
Depression	.63	—	—
Anxiety	.67	.70	—
General distress	.88	.88	.92
Prisoner			
Depression	.58	—	—
Anxiety	.78	.71	—
General distress	.86	.88	.93
Student			
Depression	.42	—	—
Anxiety	.60	.67	—
General distress	.78	.83	.87

Mokken scaling

Clinical sample and prisoner sample. For the clinical and prisoner samples, the H_i and H values obtained by applying the TEST procedure are presented in Table 1 and Table 2 (last two columns), respectively; Table 5 (upper and middle panel) presents the results for the search procedure for these samples. We first discuss the results for the clinical sample. In the clinical sample, the H_i values within the subscales ranged from .50 to .75, and for the total scale they ranged from .43 to .65. The H values for the somatization, depression, and anxiety subscales were .53, .64, and .63, respectively. The H value for the total scale equaled .53. Thus, all 18 items form a strong unidimensional scale. It is clear that both the subscales and the total scale are strong scales: H values were larger than .50. Note that H for the somatization scale was similar to the H value for the total scale. This implies that the individual somatization items related equally as well to their own subscale items as to the depression and anxiety items. The H values for the other two subscales are somewhat higher than the H value for the total scale, but differences are small.

In Table 5 (upper panel), the results of the SEARCH procedure are presented for different values of c for the clinical sample. When $c = .40$, all 18 items form a Mokken scale. For $c = .50$, 13 items form a scale that consists of mostly depression and anxiety items. For $c = .60$, only anxiety and depression items were selected for the first scale. As will be further illustrated below in the parametric analysis, this was due to the higher discrimination parameters of these items: The somatization items were less related to psychological distress than the anxiety and depression items. Furthermore, note that the scales did not comply with the theoretical dimensional structure presented in Tables 1–3; a few small scales were found and several items were excluded. Thus, the results fit the pattern one would expect when there is a unidimensional scale, as described by Sijtsma and Molenaar (2002).

For the prisoner sample a similar picture arises. Using the test procedure, it is clear that all 18 items form a strong scale ($H = .51$) and that all subscales form strong scales (see Table 2). Using the search procedure (Table 5, middle panel), again, for $c = .40$ all items form one scale and for $c = .60$ only depression and anxiety items are selected in the first scale, whereas in the second scale anxiety and somatisation items are selected.

Student sample. For the student sample a different pattern arises. The H_i values and H value for the total scale were much lower than those for the clinical and prisoner samples (Table 3, last two columns). However, for the anxiety and depression subscales the H_i values point at median scales, whereas the H_i values of the somatization scale were low, with one item (Item 4, "pains in chest") that had an H_i value lower than $H = .3$. This is reflected by the results of the search algorithm. Table 5 (lower panel) shows that even when we are liberal and accept an item cluster as unidimensional using $c = .3$, only the depression and anxiety items were selected, with one exception (Item 16). Thus, for this sample the somatization items did not form a scale and were not useful to discriminate between persons. Moreover, when these items are included in the scale, the quality of the scale is less than when these items are not included. Furthermore, for $c = .4$, five depression items formed a scale together with one anxiety item, and four anxiety items formed a scale. This may point at multidimensionality. However, increasing to $c = .5$, only three anxiety items discriminate well enough to form a scale. Thus, there are too few high-quality items to form an anxiety scale in this population

Parametric Analysis

In this parametric analysis we mainly report the results with respect to the clinical and the student sample. The results obtained from the prisoner sample were comparable with the clinical sample. For the clinical sample, in Table 6, the α_i estimates and the item location β_i parameter estimates are presented for the 18 items, as well as their corresponding theoretical dimensions. An interesting pattern in the slopes and information of the items is observed: Nine out of the 10 best discriminating items (i.e., having the highest estimated alpha values) were the depression and the anxiety items and the least discriminating anxiety item (Item 15, "feeling restless") had a clearly somatic content. These results

Table 5
Scales Obtained by Applying the SEARCH Procedure in MSP 5.0 for Different Samples

Sample	Scale 1	Scale 2	Scale 3
Clinical			
.30	1–18		
.40	1–18		
.50	2, **3**, 5, **6**, **7**, 8, **9**, 11, **12**, ***13***, 14, ***16***, 18	***1***, ***4***	
.60	2, 5, 8, **9**, 11, 14, **18**	**3, 6, 12**	
Prisoner			
.30	1–18		
.40	1–18		
.50	2, 5, **6**, 8, **9**, 11, **12**, ***13***, 14, 17, 18	1, **15, 3**	
.60	5, **6**, 8, 11, 14, **18**	**3, 9**, ***10***, **12**, ***13***	
Student			
.30	2, **3**, 5, **6**, 8, **9**, 11, **12**, 14, **15–18**	***13***, ***7***	***10***, 1
.40	2, 5, *6*, 8, 11, 14	**12, 18, 9, 3**	***13***, 7, ***16***
.50	2, 5, 8, 11, 14	**12, 18, 9**	6, ***16***
.60	2, 5, 8	11, 14	

Note. Unemphasized numbers refer to the depression items, **bold** numbers to the anxiety items, and ***bold italic*** numbers to the somatization items.

Table 6
Estimated Item Parameters (SD) for the Graded Response Model (Clinical Sample)

Item	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Som 16	2.43 (0.27)	−0.48 (0.10)	0.21 (0.10)	0.80 (0.12)	1.52 (0.16)
Som 7	1.77 (0.23)	−0.25 (0.11)	0.54 (0.13)	1.08 (0.17)	1.80 (0.26)
Som 13	1.66 (0.22)	−0.20 (0.11)	0.85 (0.17)	1.36 (0.22)	2.09 (0.30)
Som 1	1.65 (0.22)	0.37 (0.12)	0.99 (0.17)	1.62 (0.25)	2.82 (0.46)
Som 10	1.72 (0.26)	0.08 (0.12)	0.97 (0.18)	1.65 (0.24)	2.65 (0.42)
Som 4	1.60 (0.22)	−0.03 (0.12)	0.90 (0.17)	1.52 (0.23)	2.55 (0.42)
Dep 8	2.71 (0.30)	−1.18 (0.11)	−0.31 (0.09)	0.30 (0.09)	0.92 (0.13)
Dep 2	2.57 (0.28)	−0.70 (0.09)	0.22 (0.09)	0.61 (0.11)	1.21 (0.17)
Dep 5	2.54 (0.21)	−1.14 (0.16)	−0.07 (0.12)	0.57 (0.15)	1.26 (0.22)
Dep 14	2.17 (0.25)	−0.85 (0.12)	−0.01 (0.10)	0.61 (0.12)	1.25 (0.16)
Dep 11	2.01 (0.24)	−0.66 (0.11)	0.28 (0.11)	0.89 (0.14)	1.49 (0.22)
Dep 17	1.37 (0.21)	0.43 (0.15)	1.66 (0.29)	2.61 (0.42)	3.25 (0.59)
Anx 6	3.26 (0.35)	−1.25 (0.10)	−0.34 (0.07)	0.31 (0.08)	0.96 (0.13)
Anx 3	2.06 (0.22)	−1.23 (0.13)	0.08 (0.11)	0.57(0.11)	1.53 (0.19)
Anx 18	2.67 (0.29)	−0.77 (0.10)	0.09 (0.09)	0.09 (0.09)	1.39 (0.16)
Anx 12	2.46 (0.28)	−0.38 (0.09)	0.36 (0.10)	0.92 (0.14)	1.61 (0.18)
Anx 9	2.57 (0.27)	−0.54 (0.09)	0.40 (0.10)	0.90 (0.14)	1.40 (0.17)
Anx 15	1.61 (0.21)	−0.60 (0.13)	0.57 (0.14)	1.18 (0.19)	1.95 (0.29)

Note. Som = somatization; Dep = depression; Anx = anxiety.

imply that the depression and anxiety items had a stronger relationship with the general distress trait level than the somatization items. For comparison, when the items are ordered according to the item *H* values instead of the estimated alpha values, also nine out of the 10 items with the highest item *H* values are depression and anxiety items.

Because α_i is related to the item information (a larger α_i implies more information), the depression and anxiety items tend to provide more information about general distress than the somatization items. To illustrate this, consider the item information curves for three items in Figure 1. On the *x* axis the estimated latent trait values are given in standard score form. The mean of the latent trait (estimated $\theta = 0$) reflects the mean of this specific clinical population. Thus all IRT indices must be interpreted relative to the metric in the calibration sample. As can be seen in Figure 1, Item 6 from the anxiety scale ("feeling tense") provided much more information (with a peak above three) than Item 4 from the somatization scale ("pains in chest," with a peak below 1). Because the information is inversely related to the standard error of measurement, items that provide little information did not add much to the reliability of the scale.

For most items with relatively high estimated alpha parameters, the item location parameters and thus the item information was symmetrical around estimated $\theta = 0$ (see Table 6). For example, for the item "I feel blue" (Item 8), the item location ranged between –1.18 and 0.92; and for the item "feeling tense" (Item 6), the item location ranged between –1.25 and 0.96. That is, the standard error of the trait estimate is similar for persons with $\theta = -1$ and $\theta = +1$. This contrast with the findings in many clinical studies (see Reise & Waller, 2009) that the majority of items have

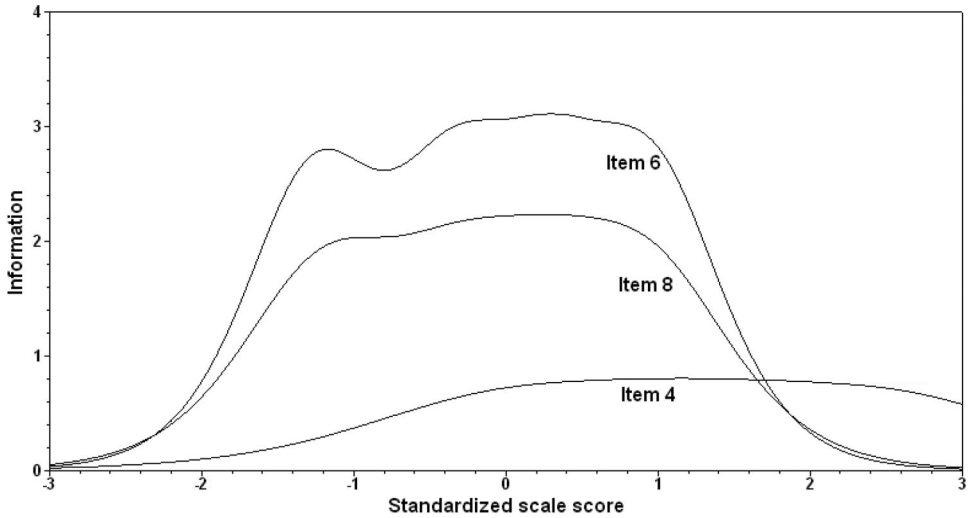

Figure 1. Distribution of item information for Items 6, 8, and 4.

peaked information for elevated trait scores. An explanation may be that items like "I feel blue" and "feeling tense" express feelings that are characteristic for the population of interest.

Figure 2 shows the distribution of information for the total scale and the distribution of the standard error of measurement in the clinical sample. The scale provided most information for estimated θ values between $\theta = -1$ and $\theta = 1$, with a maximum between 0 and 1. That is, information was highest for values of theta slightly to the right from the middle of the scale. This reflects the distribution of the scores, which were a little skewed to the right (remember that information is highest for thetas close to the item location β_i). In practice this means that the scale discriminated best between patients with average to moderately high general distress scores. To illustrate the difference with the student sample, we plotted the information for the total scale in Figure 3. Remember that in the student sample, as in the medical samples analyzed by Durá et al. (2006), Recklitis et al. (2006), and Galdón et al. (2008), mean item scores were low, indicating little distress. This has an effect on the information curves. Because item discrimination is defined as the steepness of the IRF at the item difficulty level, most items only discriminated at the higher latent trait level, which was nicely reflected by the total information function in Figure 3. The scale provided most information (measurement precision) between an estimated $\theta = 1$ and an estimated $\theta = 2.5$. Thus, within the population of students but probably also in the medical sample discussed in earlier studies, test scores are unreliable for latent trait values, say, below the mean. In these populations the scale only discriminates between persons at the very high end of the latent trait scale. This implies that when calculating change scores, for example, a researcher should be aware of the fact that the width of the confidence intervals may differ for different theta values. When using change score formulas, these different confidence intervals should be taken into account (see Reise & Havilund, 2005).

The peaked information for the different populations reflects what Reise and Waller (2009) called the "quasi-trait status" of many psychopathology constructs. The term quasi-trait means that the trait is unipolar (relevant only in one direction) and that

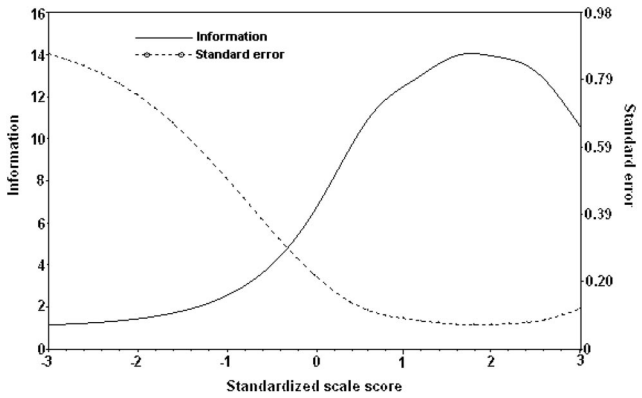

Figure 3. Distributions of total scale information and standard error of measurement in the student sample.

variation at the low end of the scale is less informative in both a substantive as well as a psychometric sense. For example, the low end of narcissism is not self-hatred but rather an absence of self-absorption, and the low end of depression is not happiness but absence of depression. The point is that the anxiety, depression, and somatization scales of the BSI-18 are undefined at the lower end of the trait.

Discussion

In this study the psychometric structure of the BSI-18 was investigated using Mokken scaling and the graded response model. For the clinical and prisoner samples, the pattern of correlations between items, subscales, and total scales, the outcome of the SEARCH procedure in MSP5.0, and the pattern of *H* values for the subscale and total scale all indicate a strong unidimensional structure underlying the items. This outcome contrasts with the dimensionality structure that has been reported in the literature thus far. Durá et al. (2006); Recklitis et al. (2006); Andreu et al. (2008), and Galdón et al. (2008) all preferred a three- or four-dimensional

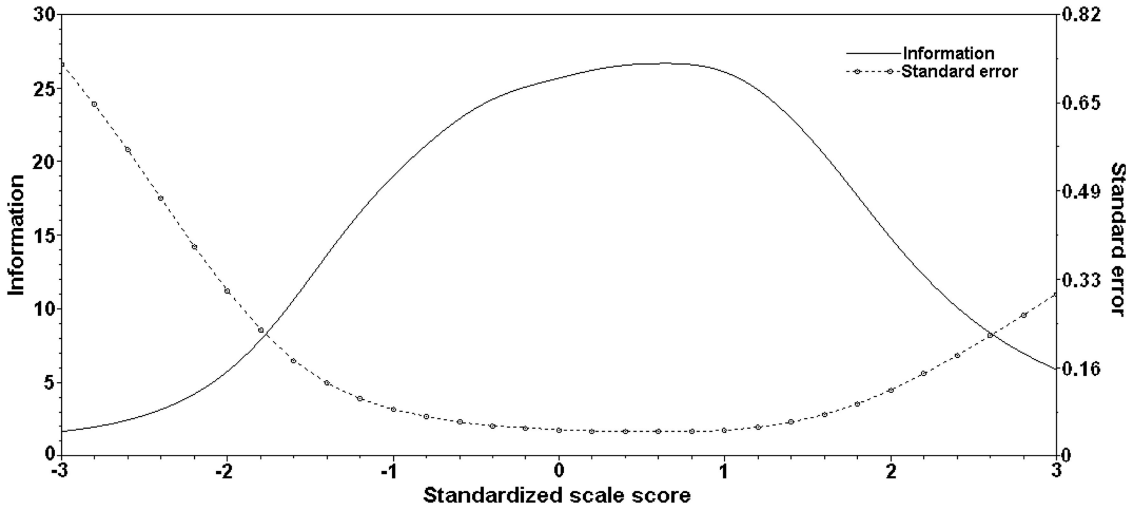

Figure 2. Distributions of total scale information and standard error of measurement in the clinical sample.

structure above a unidimensional structure based on model fit. Although differences may be due to different types of samples (cancer patients in earlier research vs. clinical, prisoner, and student samples in the present study), we think that at least part of the difference may be explained by the way results are interpreted. In general, sets of items will always show multidimensionality to some extent, and a multidimensional model will therefore always fit better than a unidimensional model. Our scale analysis, however, showed that all 18 items formed a strong Mokken scale. This is in line with earlier studies that investigated the dimensionality of the BSI consisting of 54 items (e.g., Piersma et al., 1994).

Of the four studies cited above, only two studies estimated a unidimensional model in addition to a three- and four-dimensional model (Galdón et al., 2008; Recklitis et al., 2006). Results of Recklitis et al. (2006) showed that the fit of the unidimensional model was only slightly worse than the fit of the multidimensional model, and the fit indices did not even reach the criteria levels. Galdón et al. (2008) did not find such satisfactory results for the unidimensional model, but neither did they for the three- and four-dimensional models. Only after improving the three-dimensional model by including the correlations between some errors did the fit become acceptable. However, improving the unidimensional model in this way might have resulted in a satisfactory model as well. Thus, on the basis of a slightly different interpretation of the results of previous studies and the results of the present study, we conclude that there is a strong common factor underlying all items of the BSI-18, at least for the clinical and the prisoner sample.

Another argument for why we are skeptical about forming subscales is that subscales are often so unreliable compared to composite scores that the composite scores often better predict the true score on a subscale than the subscale score itself. Sinharay, Puhan, and Haberman (2010) showed, using results from operational and simulated data, that diagnostic scores have to be based on a sufficient number of items and have to be sufficiently distinct from each other to be worth reporting and that several operationally reported subtest scores are actually not worth reporting. Emons, Sijtsma, and Meijer (2007) also showed that the classification consistency using short scales (at most 15 items) is at most 50%.

Another interesting finding in the clinical and prisoner samples was that the overall construct of psychological distress, defined by the somatization, depression, and anxiety items, is in particular defined by the depression and anxiety items. The somatization items are clearly less related to the overall construct of psychological distress. The parametric analysis showed that nine out of 10 best discriminating items all came from the depression and anxiety subscales, and most of the worst discriminating items came from the somatization subscale. For example, the estimated discrimination parameter of the anxiety item "feeling tense" was twice as high as the estimated discrimination parameter of the somatization item "pains in heart or chest." This difference was also reflected in the item information curves.

Thus, although our scale analysis showed that each of the three a priori scales is a strong scale, when the scales are combined to form one scale to measure the overall construct psychological distress, the depression and anxiety items mostly define the scale. Although theoretically psychological distress is claimed to be a combination of depression, anxiety, and somatization, it is more of a depression/anxiety scale.

Results for the student sample showed that the depression and anxiety items form one scale and that the somatization items do not discriminate between student's total scores. It is thus important to realize that when the BSI-18 is used as a screening tool in a population with a low distress level, such as in a general population or in a student population, the scale characteristics may be different from those in a population with a high distress level.

Some authors claim that although high intercorrelations may exist between several factors, as we found in the present study, the scale may still be multidimensional because factors may have a different pattern of correlations with other measures relevant for a particular patient group and that these patterns of correlations should be identified. We do not think that this is a very fruitful strategy. Any two items that are not perfectly correlated must correlate with an external variable differently. When we should follow then the strategy of looking at different correlation patterns, scale analysis would be irrelevant. Instead, we advocate a strategy where in order for a measure's external correlates to be meaningful, a coherent latent structure must be identified first. To the extent that this is not the case (such as for the somatization items in the student sample), it is challenging to fully understand the sources of the BSI-18 score variation. The strong relation between the BSI-18 scales and their theoretical structure does have practical significance; it indicates that the scales can be meaningfully interpreted in the population of prisoners and clinical patients as a general distress scale. Note that this is concordant with the original idea by Derogatis (2001) that the BSI-18 is constructed to be more homogeneous than earlier instruments. The results also emphasize that one should be careful in clinical practice to overemphasize the difference between depression, anxiety, and somatization subscale scores, because subtest scores are highly related.

We hypothesize that relations between subtest scores and other variables will not show large differences and that much of the variance explained in the subscores is due to the underlying factor of psychological distress (for a related discussion, see the remarks made by Reise & Waller, 2009, with respect to the bifactor model). Although clinicians and educational researchers use diagnostic scores on subtests as added value to the total score, we fully agree with Sinharay et al. (2010) that these scores are only useful when they provide a more accurate measure of the construct being measured (e.g., depression or algebra) than is provided by the total score (psychological distress or content knowledge of mathematics). Therefore, future research regarding convergent and discriminant validity of the BSI-18 may take the general factor of psychological distress into account (Brouwer, Meijer, Weekers, & Baneke, 2008) .

References

Andreu, Y., Galdón, M. J., Durá, E., Ferrando, M., Murgui, S., García, A., & Ibáñez, E. (2008). Psychometric properties of the Brief Symptoms Inventory–18 (BSI-18) in a Spanish sample of outpatients with psychiatric disorders. *Psicothema, 20,* 844–850.

Brouwer, D., Meijer, R. R., Weekers, A. M., & Baneke, J. J. (2008). On the dimensionality of the Dispositional Hope Scale. *Psychological Assessment, 20,* 310–315. doi:10.1037/1040-3590.20.3.310

Derogatis, L. R. (1983). *SCL–90–R administration, scoring, and proce-*

dures manual (2nd ed.). Baltimore, MD: Clinical Psychometric Research.

Derogatis, L. R. (2001). *Brief Symptom Inventory (BSI)-18: Administration, scoring and procedures manual.* Minneapolis, MN: NCS Pearson.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory (BSI): An introductory report. *Psychological Medicine, 13,* 595–605. doi:10.1017/S0033291700048017

Durá, E., Andreu, Y., Galdón, M. J., Ferrando, M., Murgui, S., Poveda, R., & Jimenez, Y. (2006). Psychological assessment of patients with temporomandipular disorders: Confirmatory analysis of the dimensional structure of the Brief Symptoms Inventory 18. *Journal of Psychosomatic Research, 60,* 365–370. doi:10.1016/j.jpsychores.2005.10.013

Egberink, I. J. L., & Meijer, R. R. (2010). An item response theory analysis of Harter's self-perception profile for children or why strong clinical scales should be distrusted. *Assessment.* Advance online publication. doi:10.1177/1073191110367778

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12,* 105–120. doi:10.1037/1082-989X.12.1.105

Emons, W. H. M., Meijer, R. R., & Denollet, J. (2007). Negative affectivity and social inhibition in cardiovascular disease: Evaluating type-D personality and its assessment using item response theory. *Journal of Psychosomatic Research, 63,* 27–39. doi:10.1016/j.jpsychores.2007.03.010

Galdón, M. J., Durá, E., Andreu, Y., Ferrando, M., Murgui, S., Pérez, S., & Ibañez, E. (2008). Psychometric properties of the Brief Symptom Inventory-18 in a Spanish breast cancer sample. *Journal of Psychosomatic Research, 65,* 533–539. doi:10.1016/j.jpsychores.2008.05.009

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Lambert, M. C., Schmitt, N., Samms-Vaughan, M. E., An, J. C., Fairclough, M., & Nutter, C. A. (2003). Is it prudent to administer all items for each child behavior checklist cross-informant syndrome? Evaluating the psychometric properties of the youth self-report dimensions with confirmatory factor analysis and item response theory. *Psychological Assessment, 15,* 550–568. doi:10.1037/1040-3590.15.4.550

Meijer, R. R. (2010). A comment on Watson, Deary, and Austin (2007) and Watson, Roberts, Gow, and Deary (2008): How to investigate whether personality items form a hierarchical scale? *Personality and Individual Differences, 48,* 502–503. doi:10.1016/j.paid.2009.11.004

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9,* 354–368. doi:10.1037/1082-989X.9.3.354

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment, 90,* 227–238. doi:10.1080/00223890701884921

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows: A program for Mokken scale analysis for polytomous items* (Version 5.0) [User's manual]. Groningen, the Netherlands: IecProGAMMA.

Moorer, P., Suurmeijer, Th. P. B. M., Foets, M., & Molenaar, I. W. (2001). Psychometric properties of the RAND-36 among three chronic diseases (multiple sclerosis, rheumatic diseases and COPD) in the Netherlands. *Quality of Life Research, 10,* 637–645. doi:10.1023/A:1013131617125

Piersma, H. L., Boes, J. L., & Reaume, W. M. (1994). Unidimensionality of the Brief Symptom Inventory (BSI) in adult and adolescent inpatients. *Journal of Personality Assessment, 63,* 338–344. doi:10.1207/s15327752jpa6302_12

Recklitis, C. J., Parsons, S. K., Shih, M., Mertens, A., Robison, L. L., & Zeltzer, L. (2006). Factor structure of the Brief Symptom Inventory-18 in adult survivors of childhood cancer: Results from the Childhood Cancer Survivor Study. *Psychological Assessment, 18,* 22–32. doi:10.1037/1040-3590.18.1.22

Reise, S. P., & Havilund, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment, 84,* 228–238. doi:10.1207/s15327752jpa8403_02

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5,* 27–48. doi:10.1146/annurev.clinpsy.032408.153553

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Iowa City, IA: Psychometric Society.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 10,* 345–359. doi:10.1037/1040-3590.10.4.345

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

Sinharay, S. Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research, 45,* 553–573. doi:10.1080/00273171.2010.483382

Teresi, J. A., Kleinman, M., Ocepek-Welikson, K., Ramirez, M., Gurland, B., Lantigua, R., & Holmes, D. (2000). Applications of item response theory to the examination of the psychometric properties and differential item functioning of the comprehensive assessment and referral evaluation dementia diagnostic scale among samples of Latino, African American, and white non-Latino elderly. *Research on Aging, 22,* 738–773. doi:10.1177/0164027500226007

Thissen, D., Chen, W. H., & Bock, R. D. (2003). MULTILOG (Version 7) [Computer software]. Lincolnwood, IL: Scientific Software International.

van Ginkel, J. R., van der Ark, A. L., & Sijtsma, K. (2007). Multiple imputation of items scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research, 42,* 387–414.

Waller, N. G. (2008). Commingled samples: A neglected source of bias in reliability analysis. *Applied Psychological Measurement, 32,* 211–223. doi:10.1177/0146621607300860

Received September 18, 2009
Revision received July 2, 2010
Accepted July 9, 2010 ■