View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DSpace at VU

Psychological Assessment
2013, Vol. 25, No. 2, 520–531

© 2013 American Psychological Association
1040-3590/13/$12.00 DOI: 10.1037/a0031669

Response Shifts in Mental Health Interventions: An Illustration of Longitudinal Measurement Invariance

Marjolein Fokkema, Niels Smits, Henk Kelderman, and Pim Cuijpers
Vrije Universiteit Amsterdam

The efficacy of treatments for depression is often measured by comparing observed total scores on self-report inventories, in both clinical practice and research. However, the occurrence of response shifts (changes in subjects' values, or their standards for measurement) may limit the validity of such comparisons. As most psychological treatments for depression are aimed at changing patients' values and frame of reference, response shifts are likely to occur over the course of such treatments. In this article, we tested whether response shifts occurred over the course of treatment in an influential randomized clinical trial. Using confirmatory factor analysis, measurement models underlying item scores on the Beck Depression Inventory (Beck & Beamesderfer, 1974) of the National Institute of Mental Health Treatment of Depression Collaborative Research Program (Elkin, Parloff, Hadley, & Autry, 1985) were analyzed. Compared with before treatment, after-treatment item scores appeared to overestimate depressive symptomatology, measurement errors were smaller, and correlations between constructs were stronger. These findings indicate a response shift, in the sense that participants seem to get better at assessing their level of depressive symptomatology. Comparing measurement models of patients receiving psychotherapy and medication suggested that the aforementioned effects were more apparent in the psychotherapy groups. Consequently, comparisons of observed total scores on self-report inventories may yield confounded measures of treatment efficacy.

Keywords: response shift, longitudinal measurement invariance, Beck Depression Inventory, reliability

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Assessment of Change

Self-report instruments like the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) are often used to measure depression in mental health practice and research. In these instances, the total score is taken as a measure of the construct of interest, in this case, depression severity. A single score may be used to measure depression severity, for example, to assess whether a patient is eligible for psychotherapy. But in many cases, scores are compared, either over time or between groups. For example, in clinical practice, treatment outcomes may be assessed at regular time intervals by means of routine outcome monitoring (ROM; e.g., De Beurs et al., 2011; Lambert, Hansen, & Finch, 2001). In ROM, the same self-report instruments are administered at fixed time intervals over the course of treatment, informing therapists and patients about increases or declines in depression scores. Group comparisons are made, for example, in a randomized clinical trial (RCT), in which the effect of therapies are assessed by calculating the difference between the mean total scores of the experimental group and the control group. Often, analysis of covariance (ANCOVA) is applied, in which posttreatment group means are compared after adjusting for differences in pretreatment means. All of these comparisons rely on observed scores, which are taken as a measure of the construct of interest. Therefore, it is assumed that raw score differences between groups, or over time, indicate differences in true scores on the construct of interest.

Self-reports, however, are subjective measures by nature. The interpretation of self-report items and response categories may vary between persons, between groups, or over time. They are not objective, directly observable outcomes like the number of cigarettes smoked, or the number of pounds gained in weight. Moreover, interpretation of self-report items may change considerably in the course of treatment. Patients' standards (e.g., what is "normal"), their understanding of which behaviors and symptoms constitute depression, or their awareness of these behaviors and symptoms may change. For instance, for many modern psychotherapies for depression, treatment guidelines stress the importance of providing the patient with psychoeducation (e.g., Beck, 1979; Klerman, Weissman, Rounsaville, & Chevron, 1984). In the initial sessions, the therapist will explain what depression is and that many of the symptoms a patient is suffering from are part of depression. This could change patients' concept of the disorder and the way they view their symptoms. In turn, this may influence the way in which patients respond to self-report items, posing a potential problem for comparing observed pre- and posttreatment scores. In contrast, such changes may not occur among patients solely taking antidepressant medication, so comparisons of observed scores between different treatment groups may be confounded as well.

This article was published Online First January 21, 2013.

Marjolein Fokkema, Niels Smits, Henk Kelderman, and Pim Cuijpers, Faculty of Psychology, Vrije Universiteit Amsterdam, the Netherlands.

Correspondence concerning this article should be addressed to Marjolein Fokkema, Faculty of Psychology, Vrije Universiteit Amsterdam, Room 2B73, Van der Boechorststraat 1, 1081BT Amsterdam, the Netherlands. E-mail: m.fokkema@vu.nl

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

To allow for valid comparisons of observed scores, it has to be ascertained that these scores represent the construct of interest in the same way, and are not confounded by response shifts: changes in subjects' standards of measurement. If observed scores are confounded, they may offer very limited insight into the true progress patients make, or the relative efficacy of treatments for depression. In clinical practice, for example, this may mean that a ROM assessment indicates a deterioration, while a patient is actually improving. Or in a clinical trial, invalid conclusions may be drawn about the relative efficacy of treatments.

Whereas the importance of response shifts has been recognized in other fields of psychology (e.g., Ahmed, Bourbeau, Maltais, & Mansour, 2009; Ahmed et al., 2005; Golembiewski, Billingsley, & Yeager, 1976; Norman & Parker, 1996; Oort, 2005; Schaubroeck & Green, 1989; Schmitt, 1982; Vandenberg & Self, 1993), the topic of response shift has, to our knowledge, been largely absent in mental health literature. Because response shifts pose a potential threat to the validity of test score comparisons, studying response shifts within a mental health context seems crucial. Therefore, the current study has two aims: first, to draw to attention the subject of response shift in mental health research and, second, to study potential response shifts in self-report data of an influential clinical trial.

In what follows, we provide an example by analysis of BDI scores from the National Institute of Mental Health (NIMH) Treatment of Depression Collaborative Research Program (TDCRP; Elkin, Parloff, Hadley, & Autry, 1985; Elkin et al., 1989). The NIMH TDCRP has been very influential in depression treatment research, and in the development of treatment guidelines. The results of the study are included in meta-analyses up to date (e.g., Cuijpers et al., 2011), and GoogleScholar reported over 2,000 citations (March 2012). The NIMH TDCRP was aimed at establishing the relative efficacy of two psycho- and two pharmacotherapies for depression: cognitive behavioral therapy (CBT), interpersonal psychotherapy (IPT), the tricyclic antidepressant Imipramine, and placebo pills. The study was very carefully designed and adequately powered (Elkin et al., 1985). With a sample size of nearly 250 subjects, it is one of the largest studies in depression treatment research. Data on the BDI are used, as it is one of the most widely used self-report instruments to measure depression (Beck, Steer, & Garbin, 1988; Shafer, 2006). To clarify the theory and rationale of the methods applied in this article, a typology of change and the related concept of measurement invariance are discussed first.

Response Shift and Measurement Invariance

In their classic article on measurement of change, Golembiewski et al. (1976) introduced a distinction between the measurement of real change and measurements that are confounded by changes in the subject's frame of reference. For example, measurements can be confounded by a recalibration of the scales used to measure a construct, or a redefinition of the construct being measured may occur (see also Oort, 2005). Whereas Golembiewski et al. (1976) used exploratory factor analysis to study shifts in subjects' frame of reference in the course of an intervention, most authors since have made use of confirmatory factor analysis (CFA; e.g., Millsap & Hartog, 1988; Oort, 2005; Schmitt, 1982).

CFA allows researchers to test whether the parameters of a hypothesized factor model are equal across populations. In a series of nested CFA models, equality restrictions on items' factor loadings, intercepts, and residual variances can be applied in order to test for their equality. This broad area of research is usually referred to as *measurement invariance* (MI; Millsap, 2011). Vandenberg and Lance (2000) provide a thorough review of practices in studying MI. In most cases, MI is studied across groups, for example, in cross-cultural research, to compare measurement models of questionnaires across countries (e.g., Zhang et al., 2011). By contrast, MI can be studied longitudinally as well, to compare measurement models in the same sample across time points (Vandenberg & Lance, 2000). Longitudinal MI can be used as a framework to study and test the occurrence of response shifts: When a response shift has occurred, measurement models are expected to differ between time points (Oort, 2005).

Although several other interpretations for differences found in factorial models across occasions have been proposed (Golembiewski et al., 1976; Millsap & Hartog, 1988; Schmitt, 1982), we limit our discussion to Oort's (2005). In this operationalization, different salient loadings across measurements indicate a redefinition, or reconceptualization, of the construct of interest. Differences in the sizes of factor loadings indicate a reprioritization, which means that some items have become more (or less) important to the measurement of the construct of interest. Differences in item intercepts indicate uniform recalibration: a recalibration of the item scale, which influences all response options within an item, and all subjects to the same extent and in the same direction. When there is a recalibration of the item scale, but this recalibration differs in extent or direction across subjects, or response options within an item, the recalibration is called *nonuniform*. Nonuniform recalibration is indicated by differences in residual variances. Finally, differences in factor correlations across time signifies a reconceptualization or reprioritization at a higher level (Oort, 2005). The procedure to test whether loadings, intercepts, and residual variances vary over measurement occasions is further explained in the Method section.

In the current study, we test the occurrence of response shifts over the course of the NIMH TDCRP by comparing the measurement models underlying pre- and posttreatment BDI scores. In addition, we compare measurement models underlying the posttreatment scores of the psycho- and pharmacotherapy groups to see whether a potential response shift can be (partly) attributed to psychotherapy. This allows us to test hypotheses on psychometric properties of an instrument commonly used in clinical practice and to formulate substantive theories about potential lacks of longitudinal invariance.

Method

Data Set

Study. Data of NIMH TDCRP was used, which has been extensively described elsewhere (Elkin et al., 1985). The study was carried out at universities in Washington, D.C., Oklahoma, and Pittsburgh. Of a total of 560 prospective subjects, 250 met study criteria and were randomly assigned to treatment conditions. After randomization, before the first treatment session, 11 patients

dropped out (five in the psychotherapy and six in the medication conditions).

Subjects. All subjects in the study were diagnosed with major depressive disorder (MDD), according to Research Diagnostic Criteria (Spitzer, Endicott, & Robins, 1978). Of the 239 patients who entered treatment, 71 (30%) were male, and 63 (26%) were single. The majority of subjects were White (212, 89%), 21 were Black (9%), and five were Hispanic (2%). One subject did not report his racial or ethnic identity. Average age was 35 years ($SD = 8.5$). Ninety-six participants (40%) were college graduates, 83 (35%) had had some college education, and 60 (25%) had high school education or less. Most participants (64%) reported to have had at least one previous episode of MDD.

Treatments. Treatments consisted of 16 weeks of CBT (Beck, 1979), IPT (Klerman et al., 1984), Imipramine (an antidepressant) plus clinical management (IMI-CM), or placebo plus clinical management (PLA-CM). All treatments consisted of 16–20 meetings. Duration of psychotherapy sessions was 50 min. Duration of CM sessions was 20–30 min, with the exception of the initial session, taking 45–60 min.

Outcome. A total of 155 participants completed treatment, resulting in a dropout of 35%. One of the main outcome analyses comprised an ANCOVA on BDI total scores, with pretreatment scores as a covariate, revealing no significant differences between treatment conditions (Elkin et al., 1989). A recovery analysis indicated no significant differences between treatments either. According to the recovery criterion of a BDI score ≤ 9 (Beck & Beamesderfer, 1974; Beck et al., 1988), 29 subjects (49%) in the CBT, 34 subjects (56%) in the IPT, 30 subjects (53%) in the IMI-CM, and 25 subjects (40%) in the PLA-CM condition were recovered at posttreatment (Elkin et al., 1989).

Instruments

The BDI (Beck & Beamesderfer, 1974; Beck et al., 1961) was one of the primary outcome measures, administered before, during, and after treatment. The BDI is a 21-item questionnaire designed to measure behavioral symptoms of depression. Items deal with symptoms of depression like mood, sense of failure, crying spells, irritability, and loss of libido. All items consist of four response options, rated 0–3, indicating increasing severity. Subjects are asked to report the response that most accurately describes their current feelings. In most cases, item scores are summed to a total score ranging from 0 to 63, with higher scores reflecting higher severity of depression. The depressive symptoms covered by the items of the BDI are presented in Table 1. Satisfactory reliability estimates for BDI total scores have been reported, with average test–retest reliability of .72 and average internal consistency of .84 (Yin & Fan, 2000).

Although, in general, a total score is used for the BDI, several different factorial structures have been reported in previous studies (Beck et al., 1988; Shafer, 2006). A two-dimensional model has been found by several authors, with a factor consisting of items covering cognitive symptoms and a factor consisting of items covering somatic symptoms (e.g., Louks, Hayne, & Smith, 1989; Steer, Iguchi, & Platt, 1992). However, many authors have found a three-factor structure to be the best fitting model, with an additional factor covering performance and motivational difficulties (e.g., Byrne & Baron, 1993; Shafer, 2006).

MI of the BDI between groups has been extensively researched. Although in most cases significant differences in the size of factor loadings were found, the factorial structure underlying the BDI scores was found to be very similar across countries, and across

Table 1
Descriptives of Pre- and Posttreatment BDI Item Scores

	Incomplete data						EM algorithm[a]			
	T1			T2			T1		T2	
Item	*N*	*M*	Var	*N*	*M*	Var	*M*	Var	*M*	Var
1. Mood	239	1.82	0.64	155	0.48	0.42	1.81	0.64	0.48	0.41
2. Pessimism	239	1.62	0.76	155	0.49	0.50	1.61	0.74	0.50	0.50
3. Sense of Failure	238	1.45	0.88	155	0.46	0.56	1.45	0.91	0.48	0.55
4. Lack of Satisfaction	239	1.86	0.55	155	0.61	0.53	1.86	0.56	0.63	0.52
5. Guilty Feeling	238	1.29	0.84	155	0.34	0.36	1.30	0.86	0.34	0.35
6. Sense of Punishment	238	1.02	1.51	155	0.26	0.42	1.02	1.51	0.27	0.42
7. Self-Hate	239	1.46	0.53	155	0.59	0.56	1.46	0.54	0.60	0.55
8. Self Accusations	239	1.40	0.69	155	0.56	0.49	1.39	0.71	0.59	0.48
9. Self Punitive Wishes	239	0.75	0.46	155	0.20	0.21	0.76	0.48	0.21	0.20
10. Crying Spells	239	1.18	0.81	155	0.24	0.37	1.18	0.80	0.26	0.37
11. Irritability	239	1.18	0.52	155	0.41	0.45	1.20	0.53	0.41	0.45
12. Social Withdrawal	239	1.36	0.66	155	0.40	0.43	1.37	0.67	0.40	0.44
13. Indecisiveness	239	1.60	0.69	155	0.36	0.43	1.61	0.69	0.37	0.43
14. Body Image	238	1.11	1.05	155	0.61	0.74	1.14	1.06	0.61	0.75
15. Work Inhibition	239	1.61	0.38	155	0.45	0.40	1.61	0.38	0.46	0.39
16. Sleep Disturbance	237	1.41	0.87	155	0.57	0.59	1.41	0.88	0.59	0.58
17. Fatigability	238	1.62	0.60	155	0.54	0.48	1.64	0.61	0.55	0.48
18. Loss of Appetite	238	0.92	0.88	155	0.18	0.29	0.92	0.89	0.19	0.29
19. Weight Loss	238	0.36	0.62	155	0.05	0.10	0.37	0.64	0.06	0.09
20. Somatic Preoccupation	238	0.50	0.45	155	0.23	0.21	0.50	0.46	0.24	0.21
21. Loss of Libido	238	1.21	1.16	155	0.50	0.72	1.20	1.15	0.54	0.73

Note. BDI = Beck Depression Inventory; EM = expectation-maximization; T1 = Time 1; T2 = Time 2; Var = Variance.
[a] Parameters estimated using an EM algorithm (Schafer, 2002).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

gender groups (Byrne & Baron, 1994; Byrne, Baron, & Balev, 1996; Byrne, Baron, & Campbell, 1993, 1994; Byrne, Baron, Larsson, & Melin, 1996; Byrne & Campbell, 1999). There has been a lack of research on longitudinal MI of the BDI. To our knowledge, only Uher et al. (2008) have studied it in a simultaneous analysis of three rating scales for depression: the Hamilton Depression Rating Scale (Hamilton, 1960), the Montgomery-Åsberg Depression Rating Scale (Montgomery & Åsberg, 1979), and the BDI. Testing the longitudinal invariance of the structure of the composite pool of items, they judged the structure to be relatively invariant, but they did report a deterioration of fit for the restriction of metric invariance for the Cognitive factor, which primarily consisted of items of the BDI.

Procedure

Estimation. CFA models with mean structure were run using LISREL 8.71 (Jöreskog & Sörbom, 1996). Estimation was performed using maximum likelihood (ML). ML was not developed for analysis of ordinal variables, as it assumes a multivariate normal distribution. Alternatively, weighted least squares or robust maximum likelihood estimation may be used. However, these methods require calculation of the asymptotic covariance matrix in LISREL (Jöreskog, 1990), resulting in listwise deletion of cases with missing values. This is a generally inappropriate approach for dealing with missing data (Graham, 2009), and would in this case result in the loss of 36% of observations. Furthermore, listwise deletion may result in biased parameter estimates when data are not missing completely at random. At the same time, ML parameter estimates have been found to be trustworthy under conditions of nonnormality (Boomsma & Hoogland, 2001). As a result, ML was used for model estimation in the current study. To take into account missing data, means and (co)variances were estimated using an expectation-maximization (EM) algorithm implemented in R (R Development Core Team, 2010; Schafer, 2002), which were subsequently used to estimate the CFA models in LISREL.

In addition to the estimates based on the EM algorithm, means, variances, and reliability estimates based on the incomplete data set are presented. Because the use of alpha as an indicator of how reliable a test score measures one construct has been criticized by several authors (e.g., Hattie, 1985; Revelle & Zinbarg, 2009; Sijtsma, 2009), McDonald's ω_h (McDonald, 1999) is presented in the Results section as well. McDonald's ω_h provides an estimate of the amount of test score variance attributable to a common factor underlying all item scores.

Tests of MI. By comparing the fit of several nested models, the tenability of equality restrictions on loadings, intercepts, and residual variances can be tested. In this study, the approach recommended by Vandenberg and Lance (2000) was followed by testing the following CFA models consecutively:

1. Configural invariance: equal factor loading patterns across occasions.
2. Metric invariance: equal factor loadings across occasions.
3. Scalar invariance: equal item intercepts across occasions.
4. Uniqueness invariance: equal residual variances across occasions.

A rejection of any of these invariance models indicates a response shift. For interpretation of a lack of invariance, the operationalization by Oort (2005) is used.

To establish a well fitting baseline model, pretreatment (Time 1) data were used as a reference point. Three different factor structures were fitted to the pretreatment data: a one-dimensional model, in which all items within the same measurement occasion load on a single factor, which is implicitly assumed by the use of a total sum score; second, a two-dimensional model, in which the first 14 items loaded on the Cognitive factor, and the seven last items loaded on the Somatic factor (Louks et al., 1989; Steer et al., 1992); third, a three-dimensional model, in which Items 1 through 10 and 14 loaded on the Cognitive factor, Items 4, 11 through 13, 15, 17, and 20 loaded on the Performance factor, and Items 16, 18, 19, and 21 loaded on the Somatic factor (Byrne et al., 1993; Byrne & Campbell, 1999). In both the two- and three-dimensional models, factor covariances were included. After obtaining a well fitting baseline model, the same models were fitted to the posttreatment (Time 2) data set as well. If the model fit for the posttreatment data set was very different from that of the pretreatment data set, this was interpreted as a lack of configural invariance.

Longitudinal invariance. The equality of the parameters of the measurement model across occasions was tested by simultaneous analysis of pre- and posttreatment data sets.

In this longitudinal model, the repeated measurements were taken into account by including residual covariances between the same items over time, following recommendations by Oort (2005) and Vandenberg and Lance (2000). For the same reason, factor covariances between time points were included in the model. For these analyses, data for all treatment groups were combined, as the measurement model is assumed to be the same in all groups before treatment, due to randomization. If no changes in measurement model occurred of time in any of the treatments groups, we expected to find the same measurement model in the complete posttreatment data set.

Between-group invariance. When full longitudinal MI could not be obtained, we proceeded in an exploratory fashion with a multigroup comparison. In the multigroup comparison, the posttreatment measurement models of the psychotherapy and pharmacotherapy groups were compared, instead of measurement occasions. In this analysis, the data for CBT and IPT groups and the IMI-CM and PLA-CM groups were combined, as the small sample sizes did not allow for separate estimation of measurement models in all four treatment groups. This allowed us to assess whether a potential response shift influenced treatment groups to a different extent.

Model modifications. Modification indices and standardized residuals in LISREL output were studied for improving inadequate model fit, in two ways. First, the baseline model was refined by allowing item residuals to be correlated within time points, for a number of items. Second, when full metric, scalar, or uniqueness invariance could not be obtained, attempts were made to find a well fitting model of partial invariance (Byrne, Shavelson, & Muthén, 1989; Yoon & Millsap, 2007). Because such post hoc model modifications are susceptible to chance capitalization (MacCallum, Roznowski, & Necowitz, 1992), to minimize this risk, model modifications were made only when they made sense from a substantive point of view. Only fixed or restricted parameters with modification indices ≥ 5 and/or standardized residuals ≥ 2 were considered to be released, as suggested by Jöreskog and Sörbom (1996). All parameter restrictions were lifted one at a time.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Assessment of model fit. To evaluate model fit, the Minimum Fit Function chi-square and associated degrees of freedom were used to test the null hypothesis that the difference between the population and model parameters equals zero. Because chi-square values tend to increase with sample size and model complexity (Jöreskog & Sörbom, 1996), significant chi-square values are to be expected for any baseline model incorporating 21 observed variables. However, irrespective of the fit of the initial model, nested models can be compared by means of $\Delta\chi^2$ and Δdf, with a significant $\Delta\chi^2$ indicating a significant deterioration of fit (Jöreskog & Sörbom, 1996; Steiger, Shapiro, & Browne, 1985).

As recommended by many authors, chi-square criteria were used in conjunction with several other model fit indices (e.g., Chen, Curran, Bollen, Kirby, & Paxton, 2008; Cheung & Rensvold, 2002; Hu & Bentler, 1998). We used the following indices:

First, the comparative fit index (CFI; Bentler, 1990) compares the fit of the fitted model with the fit of an independence model, assuming no relationship between the variables. As a rule of thumb, CFI values of $\geq .90$ represent acceptable model fit, which was used as a cutoff in the current study. To assess whether there is a substantial difference in fit between two nested models, the difference in CFI values, ΔCFI, can be used. In the current study, Mead, Johnson, and Braddy's (2008) suggested cutoff of .002 was used.

Second, Aikaike's information criterion (AIC; Akaike, 1987) provides a weighted index of model accuracy and complexity. There is no fixed cutoff value for this criterion, but it can be used to compare models for the same data set: The model with the lowest AIC value is preferred.

The third index is the standardized root-mean-square residual (SRMR). Hu and Bentler (1999) recommended a cutoff value of $\leq$ .08 for adequately fitting models, which was used as a cutoff value in this study.

The fourth index is the root-mean-square rrror of approximation (RMSEA). According to Browne and Cudeck (1993), RMSEA values of $\leq .05$ represent good fit, and values $\leq .08$ represent adequate fit. In addition to the point estimates for RMSEA, 90% confidence intervals are presented as well. However, greater emphasis is placed on the SRMR, as the RMSEA is less preferable with small sample sizes ($N < 250$; Hu & Bentler, 1999).

Note that the cutoffs for RMSEA and CFI used in the current study are more lenient than the widely accepted cutoffs proposed by Hu and Bentler (1999). As noted by Marsh, Hau, and Wen (2004), these criteria may be too restrictive for item-level analyses of multifactor instruments. Bollen (1989, as cited in Marsh et al., 2004) suggested that the value of fit indices reported in practice may be taken into account when evaluating model fit. CFI values failed to exceed the .95 cutoff proposed by Hu and Bentler (1999) in earlier studies on MI of the BDI (Byrne, 1995; Byrne & Baron, 1993, 1994; Byrne et al., 1994). Consequently, a cutoff of .90 was used for CFI in the current study. Earlier studies on MI of the BDI did not report unscaled values for RMSEA.

Results

Descriptives of BDI Scores

Descriptives of item scores are presented in Table 1. All item score means show a decrease, indicating that, on average, patients improved over the course of treatment. The majority of item variances decreased (see Table 1), whereas the sum score variance increased from 60.44 before treatment to 76.96 after treatment, based on the estimates of the EM algorithm. The increased sum score variance indicates that heterogeneity has increased over the course of treatment and that some participants may have benefitted more from treatment than others. Similarly, reliability estimates for the test score increased from pre- to posttreatment: Cronbach's α was .78 at the pretreatment assessment (.70, .51, and .37 for the Cognitive, Performance, and Somatic subscales, respectively) and .92 at the posttreatment assessment (.86, .73, and .45 for the Cognitive, Performance, and Somatic subscales, respectively). McDonald's ω_h was .79 for the pretreatment assessment (.77, .64, and .59 for the Cognitive, Performance, and Somatic subscales, respectively) and .93 for the posttreatment assessment (.93, .85, and .68 for the Cognitive, Performance, and Somatic subscales, respectively). These estimates indicate that the influence of measurement error on test scores is smaller after treatment, compared with before treatment. Analysis of the incomplete data set yielded very similar results.

Dimensionality of BDI Scores

The one-, two-, and three-dimensional models were fitted to the pretreatment data set. The resulting fit indices are presented in the upper portion of Table 2, and estimates for the factor loadings are presented in Table 3. The one-dimensional model

Table 2
Model Fit Indices for Pre- and Posttreatment Data Sets

Time	Model	*df*	χ^2	CFI	SRMR	RMSEA	90% CI for RMSEA	AIC
T1	1D	189	441.46	.816	.0795	.0816	[0.0728, 0.0906]	614.66
	2D	188	387.85	.855	.0766	.0695	[0.0602, 0.0788]	532.21
	3D	186	335.18	.891	.0717	.0574	[0.0473, 0.0674]	464.08
	3D[a]	184	317.49	.903	.0684	.0539	[0.0434, 0.0641]	447.10
T2	1D	189	735.94	.928	.0735	.1148	[0.1066, 0.1232]	908.21
	2D	188	692.89	.933	.0723	.1072	[0.0988, 0.1157]	830.09
	3D	186	604.15	.945	.0674	.0987	[0.0902, 0.1070]	749.65
	3D[a]	184	550.13	.952	.0625	.0935	[0.0848, 0.1020]	702.75

Note. CFI = comparative fit index; SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation; CI = confidence interval; AIC = Akaike's information criterion; T1 = Time 1; 1D = one dimensional; 2D = two dimensional; 3D = three dimensional; T2 = Time 2.
[a] Model included residual covariances for Items 21 and 12 and Items 19 and 18.

Table 3
Estimated Factor Loadings for the Hypothesized Models for Pre- and Posttreatment Data Sets

Single-factor model			Two-factor model			Three-factor model		
Item	T1	T2	Item	T1	T2	Item	T1	T2
1	1.00	1.00	1	1.00	1.00	1	1.00	1.00
2	1.27 (0.25)	1.06 (0.08)	2	1.41 (0.29)	1.06 (0.08)	2	1.48 (0.33)	1.06 (0.08)
3	1.68 (0.30)	1.13 (0.08)	3	1.93 (0.37)	1.15 (0.08)	3	2.13 (0.43)	1.17 (0.08)
4	0.67 (0.18)	0.99 (0.09)	4	0.65 (0.20)	0.97 (0.09)	5	1.73 (0.37)	0.81 (0.07)
5	1.39 (0.27)	0.79 (0.07)	5	1.59 (0.32)	0.80 (0.07)	6	1.53 (0.39)	0.83 (0.08)
6	1.38 (0.32)	0.84 (0.08)	6	1.45 (0.36)	0.84 (0.08)	7	1.49 (0.31)	1.24 (0.08)
7	1.14 (0.22)	1.21 (0.08)	7	1.31 (0.26)	1.21 (0.08)	8	1.73 (0.36)	1.13 (0.08)
8	1.37 (0.25)	1.10 (0.08)	8	1.55 (0.30)	1.11 (0.08)	9	0.69 (0.20)	0.61 (0.05)
9	0.57 (0.16)	0.59 (0.05)	9	0.65 (0.18)	0.59 (0.05)	10	0.97 (0.27)	0.59 (0.08)
10	0.97 (0.23)	0.63 (0.08)	10	0.99 (0.25)	0.61 (0.08)	14	1.73 (0.38)	1.10 (0.10)
11	0.75 (0.18)	0.56 (0.08)	11	0.70 (0.19)	0.54 (0.08)	4	1.00	1.00
12	0.81 (0.20)	0.72 (0.08)	12	0.84 (0.22)	0.71 (0.08)	11	1.11 (0.27)	0.62 (0.08)
13	0.80 (0.20)	0.79 (0.08)	13	0.84 (0.23)	0.78 (0.08)	12	1.00 (0.28)	0.69 (0.08)
14	1.45 (0.29)	1.10 (0.10)	14	1.59 (0.34)	1.10 (0.10)	13	1.01 (0.28)	0.83 (0.08)
15	0.60 (0.15)	0.84 (0.07)	15	1.00	1.00	15	0.96 (0.24)	0.90 (0.07)
16	0.33 (0.20)	0.68 (0.10)	16	1.17 (0.41)	0.91 (0.11)	17	1.70 (0.37)	0.84 (0.08)
17	1.01 (0.21)	0.82 (0.08)	17	2.10 (0.52)	1.04 (0.10)	20	0.50 (0.20)	0.16 (0.06)
18	0.83 (0.23)	0.33 (0.07)	18	2.35 (0.59)	0.46 (0.08)	16	1.00	1.00
19	0.37 (0.17)	0.21 (0.04)	19	1.58 (0.43)	0.28 (0.05)	18	1.60 (0.55)	0.50 (0.10)
20	0.42 (0.15)	0.17 (0.06)	20	0.88 (0.30)	0.20 (0.07)	19	0.80 (0.34)	0.27 (0.06)
21	0.75 (0.25)	0.81 (0.11)	21	1.54 (0.50)	0.88 (0.12)	21	1.32 (0.49)	1.02 (0.17)

Note. T1 = Time 1; T2 = Time 2. Parenthetical values are standard errors. Standard errors are not calculated for loadings fixed to one for identification.

proved to have inadequate fit, judging by CFI and RMSEA. The two-dimensional model proved to have better, but still inadequate fit, judging by the CFI value. The three-dimensional model showed the best fit to the data judging by all fit indices, though the CFI value indicated inadequate fit. To improve model fit, residual covariances for Items 12 (Social Withdrawal) and 21 (Loss of Libido) and Items 18 (Loss of Appetite) and 19 (Weight Loss) were included in the model. Modification indices and standardized residuals for these parameters exceeded cutoffs, and it made sense from a substantive point of view, as the item contents overlapped: Items 12 and 21 relate to interest in other people and sex, and Items 18 and 19 relate to eating behavior. In addition, the correlated error between Items 18 and 19 was included by Byrne and Campbell (1999) as well. Including residual variances for these item pairs in the model resulted in an adequate fit for the adjusted three-dimensional model to the pretreatment data, judging by CFI, SRMR, RMSEA, and the best fitting model according to AIC. The same pattern was found for the posttreatment data set (see Table 2, lower portion), for which the adjusted three-dimensional model showed the best fit as well.

Longitudinal MI Analysis

Tests of invariance. The adjusted three-dimensional model was fitted to pre- and posttreatment data sets simultaneously to test for configural invariance. For this longitudinal model, configural invariance could be obtained, judging by all fit indices (see Table 4). All parameters in this model proved meaningful: All variances and factor loadings had positive values, and all factor loadings were significantly different from zero ($p < .05$).

Table 4
Model Fit Indices for MI Restrictions (Longitudinal Analysis)

Invariance model	*df*	χ^2	Δdf	$\Delta\chi^2$	p^a	CFI	ΔCFI	SRMR	RMSEA	90% CI	AIC
Configural	779	1739.56	—	—	—	.90577	—	.0720	.0666	[.0620, .0712]	1933.40
Metric	797	1778.11	18	38.54	.003	.90375	−.00202	.0756	.0668	[.0622, .0713]	1938.40
Partial metric[b]	795	1761.47	16	21.91	.146	.9052	−.0006	.0746	.0662	[.0616, .0708]	1924.16
Scalar[b]	813	1889.29	18	127.82	<.001	.8944	−.0108	.0854	.0695	[.0650, .0740]	2010.79
Partial scalar[c]	805	1778.01	10	16.92	.076	.9045	−.0007	.0760	.0660	[.0614, .0706]	1918.98
Uniqueness[c]	826	2812.94	21	1034.93	<.001	.8051	−.0994	.1050	.0889	[.0848, .0931]	2619.40
Partial uniqueness[d]	807	1794.00	2	15.99	<.001	.9032	−.0013	.0766	.0661	[.0616, .0707]	1923.01

Note. MI = measurement invariance; CFI = comparative fit index; SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation; CI = confidence interval; AIC = Akaike's information criterion. Dashes indicate values that cannot be calculated for these baseline models.
[a] *p* values for $\Delta\chi^2(\Delta df)$. [b] Unequal loadings for Items 17 and 18. [c] In addition, unequal intercepts for Items 14, 7, 8, 3, 9, 17, 6, and 13. [d] In addition, unequal uniquenesses for all items, except for 11 and 16.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Constraining factor loadings to be equal across occasions resulted in some deterioration of fit (see Table 4). The difference in chi-square value was significant, and the ΔCFI exceeded the cutoff value. By lifting the equality restriction on two loadings (one within the Performance and one within the Somatic factor), a partially metric invariant model could be obtained, which showed adequate fit according to all indices.

Constraining item intercepts of the partial metric invariant model to be equal across occasions resulted in a considerable deterioration of fit (see Table 4). $\Delta\chi^2$ was significant, ΔCFI exceeded the cutoff value, and CFI and SRMR indicated inadequate model fit. Equality restrictions of eight item intercepts (six within the Cognitive and two within the Performance factor) needed to be lifted to obtain adequate model fit.

Subsequently, item uniquenesses were constrained to be equal across occasions. This resulted in a considerable deterioration of fit, judging by all fit indices (see Table 4). All but two of the items showed initial modification indices > 5. Releasing equality restrictions on all item uniquenesses, except for these two items, still resulted in a significant deterioration of model fit, judging by $\Delta\chi^2$ (see Table 4).

Interpretation. The test of configural invariance indicated that the number and content of the concepts constituting depression remained the same, over the course of treatment. Similarly, the test of metric invariance indicated that the factor loadings remained largely the same, too. The tests of scalar invariance indicated considerable changes in item intercepts. Because the number of modifications needed to obtain a model of partial scalar invariance was quite large, parameter estimates of the partially metric invariant model are discussed and presented in Table 5.

The two noninvariant loadings were smaller after treatment, than before treatment, indicating that the items dealing with fatigability and appetite have become less indicative of the Performance and Somatic factors, respectively. The majority of item intercepts were higher after treatment, indicating a recalibration: After treatment, factor scores are overestimated by item scores, relative to the pretreatment measurements. The only notable exception to this increase was Item 13, dealing with indecisiveness, showing a lower intercept after treatment. As intercept differences were concentrated within the Cognitive factor, this subscale may be affected most by the recalibration. All items showed smaller residual variances after treatment, indicating that measurement errors have become smaller for all items. For the metric invariant items, this means that items have become better indicators of latent traits, after treatment.

Factor (co)variances and correlations of the partial metric invariant model are presented in Table 6. All factor variances showed an increase, suggesting increased heterogeneity among the participants, after treatment. Furthermore, only the Cognitive factor showed a notable association between time points; all other factor correlations were low. Within time points, factor correlations showed a stronger relationship between latent variables after treatment, which may be interpreted as a reconceptualization: Participants seem to view depression as a more unified concept after treatment.

Exploratory Analysis: Multigroup Comparisons

To further explore whether the differences observed in the longitudinal analysis affected treatment groups to the same extent, an exploratory multigroup analysis of MI was performed. In this multigroup analysis, the measurement models underlying the data sets of the subjects receiving psychotherapy and the subjects taking medication were compared.

Analysis of pretreatment data sets. A multigroup analysis was performed on the pretreatment data sets of both groups, first.

Table 5
Parameter Estimates for the Partial Metric Invariant Model of the Longitudinal Analysis

		Loadings		Intercepts		Uniquenesses	
Factor	Item	T1	T2	T1	T2	T1	T2
1	1	1.00	1.00	0.00	0.00 (−)	0.49 (.05)	0.16 (.02)
	2	1.10 (.08)	1.10 (.08)	−0.31 (.15)	−0.02 (.05)	0.54 (.05)	0.20 (.02)
	3	1.23 (.08)	1.23 (.08)	−0.72 (.16)	−0.10 (.06)	0.51 (.05)	0.19 (.02)
	5	0.87 (.07)	0.87 (.07)	−0.26 (.14)	−0.08 (.05)	0.64 (.06)	0.18 (.02)
	6	0.87 (.08)	0.87 (.08)	−0.53 (.16)	−0.15 (.05)	1.29 (.12)	0.24 (.02)
	7	1.21 (.08)	1.21 (.08)	−0.65 (.15)	0.02 (.05)	0.35 (.04)	0.16 (.02)
	8	1.15 (.08)	1.15 (.07)	−0.55 (.15)	0.04 (.05)	0.44 (.04)	0.15 (.02)
	9	0.60 (.05)	0.60 (.05)	−0.42 (.10)	−0.08 (.04)	0.34 (.03)	0.11 (.01)
	10	0.66 (.07)	0.66 (.07)	−0.05 (.14)	−0.06 (.05)	0.71 (.07)	0.28 (.03)
	14	1.16 (.10)	1.16 (.10)	−0.88 (.19)	0.05 (.07)	0.76 (.07)	0.41 (.04)
2	4	1.00	1.00	0.00	0.00 (−)	0.49 (.05)	0.21 (.02)
	11	0.64 (.07)	0.64 (.07)	−0.05 (.14)	0.01 (.06)	0.39 (.04)	0.34 (.03)
	12	0.72 (.07)	0.72 (.07)	−0.03 (.14)	−0.06 (.06)	0.56 (.05)	0.27 (.03)
	13	0.83 (.07)	0.83 (.07)	0.12 (.14)	−0.15 (.06)	0.54 (.05)	0.21 (.02)
	17	1.41 (.19)	0.82 (.08)	−1.01 (.37)	0.03 (.06)	0.41 (.05)	0.27 (.03)
	20	0.22 (.05)	0.22 (.05)	0.12 (.10)	0.10 (.04)	0.45 (.04)	0.20 (.02)
	15	0.85 (.06)	0.85 (.06)	0.02 (.13)	−0.08 (.05)	0.27 (.03)	0.15 (.02)
3	16	1.00	1.00	0.00	0.00 (−)	0.73 (.08)	0.39 (.04)
	18	1.30 (.30)	0.62 (.11)	−0.98 (.44)	−0.18 (.08)	0.58 (.08)	0.22 (.02)
	19	0.29 (.06)	0.29 (.06)	−0.06 (.10)	−0.11 (.04)	0.55 (.05)	0.08 (.01)
	21	1.12 (.16)	1.12 (.16)	−0.29 (.24)	−0.13 (.11)	0.99 (.10)	0.49 (.05)

Note. T1 = Time 1; T2 = Time 2. Parenthetical values are standard errors. Standard errors are not calculated for parameters fixed to zero or one for identification.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 6
Covariance Matrix of Latent Variables for the Partial Metric Invariance Model of the Longitudinal Analysis

	Variable	T1 Cognitive	T1 Performance	T1 Somatic	T2 Cognitive	T2 Performance	T2 Somatic
T1	Cognitive	0.188					
	Performance	0.103 (.642)	0.135				
	Somatic	0.050 (.323)	0.085 (.638)	0.129			
T2	Cognitive	0.098 (.453)	0.023 (.128)	0.003 (.017)	0.247		
	Performance	0.027 (.110)	0.028 (.136)	−0.003 (−.015)	0.225 (.812)	0.311	
	Somatic	0.036 (.202)	0.003 (.023)	−0.012 (−.079)	0.151 (.735)	0.181 (.785)	0.171

Note. T1 = Time 1; T2 = Time 2. Parenthetical values are correlations.

In an RCT, treatment groups are assumed to be equal before treatment, so any preexisting differences between the two groups can be assumed to reflect random fluctuations. Therefore, fitting restrictions of MI to the pretreatment data sets provides a benchmark for evaluating the relevance of differences found between posttreatment data sets. In the multigroup comparison, the model fitted the pretreatment data well, judging by all fit indices (see Table 7, upper portion). Applying equality restrictions on loadings, intercepts and residual variances did not influence model fit much. Small decreases in model fit could easily be resolved by lifting equality restrictions on one loading and one intercept.

Analysis of posttreatment data sets. A multigroup analysis was performed on the posttreatment data sets of both groups to explore whether differences between the treatment groups arose in the course of treatment. The multigroup configural invariant model did not fit the posttreatment data well, judging by all fit indices (see Table 7, lower portion).

Applying the restriction of metric invariance to the model significantly deteriorated model fit, judging by both $\Delta\chi^2$ and ΔCFI (see Table 7, lower portion). Releasing equality restrictions on loadings of two items (one within the Cognitive and one within the Somatic factor) resulted in a nonsignificant deterioration of model fit, compared with the model of configural invariance.

The restriction of scalar invariance significantly deteriorated model fit as well, and lifting equality restrictions for the intercepts of the same items resulted in a nonsignificant deterioration of model fit (see Table 7, lower portion).

Applying the restriction of uniqueness invariance to the partial scalar invariance model resulted in a substantial deterioration of model fit (see Table 7, lower portion). Eight equality restrictions on item uniquenesses were lifted to obtain nonsignificant $\Delta\chi^2$ and ΔCFI values: four items within the Cognitive, two items within the Performance, and two items within the Somatic factor.

Interpretation. Analysis of posttreatment data sets showed a number of differences in the measurement models of the psychotherapy and medication groups, after treatment. Because the number of lifted restrictions to obtain partial invariance were large for the uniqueness invariant model only, parameter estimates of the partial scalar invariance model are discussed and presented in Table 8.

The items for which equality restrictions on loadings and intercepts were lifted did not show the same pattern of freely estimated parameters. The item dealing with self-accusations was more indicative for the Cognitive factor and showed a higher intercept in the medication groups, whereas the item dealing with weight loss has become more indicative for the Somatic factor and had a lower

Table 7
Model Fit Indices for the Multigroup Analysis of Pretreatment (Upper Portion) and Posttreatment (Lower Portion) Data Sets

Model	*df*	χ^2	Δdf	$\Delta\chi^2$	*p*[a]	CFI	ΔCFI	SRMR[b]	SRMR[c]	RMSEA	90% CI	AIC
Pretreatment												
Configural	368	473.52	—	—	—	.9235	—	.0778	.0785	.0403	[0.0226, 0.0540]	710.69
Metric[d]	386	495.23	18	21.71	.245	.9208	−.0027	.0847	.0844	.0404	[0.0233, 0.0538]	696.55
Scalar[e]	404	522.24	18	27.01	.079	.9143	−.0065	.0846	.0839	.0412	[0.0251, 0.0543]	685.37
Uniqueness	425	536.22	21	13.98	.870	.9194	.0051	.0836	.0867	.0390	[0.0221, 0.0521]	659.52
Postreatment												
Configural	368	1181.03	—	—	—	.8930	—	.0822	.0976	.1257	[0.1169, 0.1346]	1329.02
Metric	386	1238.58	18	57.55	<.001	.8878	−.0052	.1093	.1086	.1249	[0.1163, 0.1336]	1335.66
Partial metric[f]	384	1204.16	16	23.13	.110	.8920	−.0010	.0872	.1032	.1230	[0.1143, 0.1318]	1312.51
Scalar[g]	402	1248.46	18	44.30	<.001	.8886	−.0034	.0935	.1031	.1209	[0.1124, 0.1295]	1302.29
Partial scalar[g]	400	1229.81	16	25.65	.059	.8907	−.0013	.0871	.1026	.1199	[0.1114, 0.1286]	1289.77
Uniqueness[g]	421	1396.66	21	166.85	<.001	.8715	−.0192	.1391	.1445	.1270	[0.1188, 0.1353]	1391.98
Partial uniqueness[h]	413	1247.59	13	17.78	.166	.8901	−.0006	.0861	.1074	.1183	[0.1098, 0.1268]	1279.80

Note. CFI = comparative fit index; SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation; CI = confidence interval; AIC = Akaike's information criterion. Dashes indicate values that cannot be calculated for these baseline models.
[a] *p* values for $\Delta\chi^2(\Delta df)$. [b] SRMR for psychotherapy group. [c] SRMR for medication group. [d] Lifting the equality restriction on loadings of Item 6 resulted in ΔCFI ≤ .002. [e] Lifting the equality restriction on intercepts of Item 18 resulted in ΔCFI ≤ .002. [f] Equality restriction for loadings of Items 8 and 19 lifted. [g] In addition, equality restriction for intercepts of Items 8 and 19 lifted. [h] In addition, equality restriction for uniquenesses of Items 18, 11, 3, 8, 2, 19, 1. and 4 lifted.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 8
Parameter Estimates for the Partial Scalar Invariant Model of the Multigroup Analysis of Posttreatment Data Sets

		Loadings		Intercepts		Uniquenesses	
Factor	Item	PSY	MED	PSY	MED	PSY	MED
1	1	1.00	1.00	0.00	0.00	0.12 (0.02)	0.21 (0.03)
	2	1.00 (0.07)	1.00 (0.07)	0.02 (0.05)	0.02 (0.05)	0.15 (0.02)	0.31 (0.04)
	3	1.14 (0.07)	1.14 (0.07)	−0.06 (0.05)	−0.06 (0.05)	0.11 (0.02)	0.30 (0.04)
	5	0.78 (0.07)	0.78 (0.07)	−0.03 (0.05)	−0.03 (0.05)	0.18 (0.03)	0.20 (0.03)
	6	0.85 (0.07)	0.85 (0.07)	−0.11 (0.05)	−0.11 (0.05)	0.22 (0.03)	0.26 (0.04)
	7	1.25 (0.08)	1.25 (0.08)	0.01 (0.06)	0.01 (0.06)	0.16 (0.03)	0.16 (0.02)
	8	0.95 (0.08)	1.48 (0.11)	0.13 (0.06)	−0.15 (0.07)	0.19 (0.03)	0.08 (0.02)
	9	0.59 (0.05)	0.59 (0.05)	−0.07 (0.04)	−0.07 (0.04)	0.10 (0.01)	0.13 (0.02)
	10	0.55 (0.07)	0.55 (0.07)	−0.02 (0.05)	−0.02 (0.05)	0.25 (0.03)	0.39 (0.05)
	14	1.07 (0.10)	1.07 (0.10)	0.07 (0.07)	0.07 (0.07)	0.45 (0.06)	0.43 (0.06)
2	4	1.00	1.00	0.00	0.00	0.27 (0.04)	0.15 (0.03)
	11	0.59 (0.07)	0.59 (0.07)	0.03 (0.06)	0.03 (0.06)	0.21 (0.03)	0.53 (0.07)
	12	0.68 (0.07)	0.68 (0.07)	−0.03 (0.06)	−0.03 (0.06)	0.24 (0.03)	0.35 (0.05)
	13	0.81 (0.07)	0.81 (0.07)	−0.17 (0.06)	−0.17 (0.06)	0.22 (0.03)	0.19 (0.03)
	15	0.87 (0.07)	0.87 (0.07)	−0.10 (0.06)	−0.10 (0.06)	0.16 (0.03)	0.15 (0.02)
	17	0.80 (0.08)	0.80 (0.08)	0.02 (0.07)	0.02 (0.07)	0.24 (0.04)	0.36 (0.05)
	20	0.14 (0.05)	0.14 (0.05)	0.12 (0.05)	0.12 (0.05)	0.21 (0.03)	0.21 (0.03)
3	16	1.00	1.00	0.00	0.00	0.37 (0.06)	0.45 (0.07)
	18	0.61 (0.11)	0.61 (0.11)	−0.22 (0.07)	−0.22 (0.07)	0.08 (0.01)	0.51 (0.07)
	19	0.49 (0.11)	0.10 (0.06)	−0.21 (0.07)	−0.05 (0.04)	0.10 (0.02)	0.06 (0.01)
	21	1.11 (0.19)	1.11 (0.19)	−0.16 (0.13)	−0.16 (0.13)	0.48 (0.07)	0.53 (0.09)

Note. PSY = psychotherapy group; MED = medication group. Parenthetical values are standard errors. Standard errors are not calculated for parameters fixed to zero or one for identification.

intercept in the psychotherapy groups. These differences are minor and would cancel out in observed test scores.

The differences between item residual variances were more substantial. In the partial scalar invariance model, 10 items showed appreciably larger values in the medication groups, whereas only two items showed larger values in the psychotherapy groups. These differences were distributed equally across the three factors. The nine remaining residual variances showed only small differences ($< .05$) between the two groups.

In Table 9, the factor (co)variances and correlations of the partial scalar invariant model are presented. The (co)variances involving the Cognitive factor are about twice as large in the psychotherapy as in the medication groups. In addition, the correlation between the Cognitive and the other two factors is somewhat higher in the psychotherapy groups as well.

These findings indicate that the changes found in the longitudinal model occurred to a greater extent in the psychotherapy groups. Although the differences in loadings and intercepts between the treatment groups were minor, residual variances were evidently smaller and factor correlations somewhat larger in the psychotherapy groups, compared with the groups receiving medication.

Table 9
Covariance Matrices of Latent Variables for the Partial Scalar Invariant Model of the Multigroup Analysis of Posttreatment Data Sets

Group	Factor	Cognitive	Performance	Somatic
PSY	Cognitive	0.367		
	Performance	0.301 (.821)	0.367	
	Somatic	0.200 (.821)	0.176 (.723)	0.162
MED	Cognitive	0.179		
	Performance	0.181 (.758)	0.319	
	Somatic	0.118 (.645)	0.178 (.724)	0.189

Note. PSY = psychotherapy group; MED = medication group. Parenthetical values are correlations.

Discussion

Summary and Interpretation

In the current study, we investigated whether a response shift has taken place over the course of an RCT of treatments for depression, using a longitudinal CFA model. The longitudinal model showed clear signs of response shifts, and parameter changes were in the same direction for most items. According to the test of metric invariance, two factor loadings were significantly smaller after treatment, relative to the referent items. This minor lack of metric invariance indicates that the relative importance of the items remained largely the same over treatment. The lack of scalar invariance was much more substantial, suggesting uniform recalibration for eight items (Oort, 2005). This means that posttreatment, observed item scores may overestimate the factor scores, compared with pretreatment. According to the test of uniqueness invariance, all residual variances decreased significantly from pre- to posttreatment. This indicates nonuniform recalibration, which may signify that the distance between some of the response options, as judged by the participants, have changed, but it may also indicate that not all participants have been influenced to the same extent by the response shift (Oort, 2005).

The increase in factor correlations over the course of treatment indicates an increased coherence of the latent constructs constituting depression.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Because a lack of MI was found in the longitudinal model, we investigated whether posttreatment data sets of the groups receiving psychotherapy and groups taking pills showed differences in measurement model to see whether the differences could be linked to treatments. The tests of metric and scalar invariance indicated minor differences between the treatment groups. However, the residual variances and factor correlations indicated that the response shift observed in the longitudinal model presented itself more strongly in the psychotherapy groups than in the medication groups.

These findings indicate that participants get better at assessing their levels of depressive symptomatology by means of the BDI, as the majority of factor loadings remained the same, but residual variances decreased. The meaning of the items and response options may have become more clear to the subjects, as they have been educated about depression during treatment. In addition, the strengthening of the relationships between constructs suggests that subjects view depression as a more unified concept after treatment. Finally, the significant increases in some of the item intercepts may indicate that subjects have become more aware of their depressive symptoms.

The aforementioned effects may be the result of psychological treatments for depression, as the psychotherapy groups seem to show this response shift to a larger extent. At the same time, the difference between the medication and psychotherapy groups posttreatment is not as clear and pronounced as the difference between pre- and posttreatment data sets. This may be due to the addition of clinical management to the medication conditions, consisting of weekly half-hour meetings with a psychiatrist. Undoubtedly, subjects have received some psychoeducation during these sessions, which may have elicited changes in standards in these groups as well.

Comparability of Observed Scores

One of the primary reasons for testing MI is to test whether differences in observed scores between groups or measurement occasions provide unambiguous measures of differences in the construct of interest (Horn & McArdle, 1992; Vandenberg & Lance, 2000). For meaningful comparison of observed group means, full scalar invariance is a prerequisite (Meredith, 1993). In both the longitudinal and multigroup analysis, full scalar invariance could not be obtained.

According to some authors, full MI is too strict a criterion for comparability of group means (Byrne et al., 1989; Steenkamp & Baumgartner, 1998). They argue that, in case of partial MI, group means can still be meaningfully compared by means of latent variable methodologies, as long as one item within each factor is invariant, besides the reference item. However, some authors argue that a majority of items within a factor should be invariant for meaningful comparisons (Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000). In any case, our findings indicate a substantial lack of scalar invariance in the longitudinal model, which should be taken into account by using latent variable methodologies when group means are compared over time. Comparisons involving pre- and posttreatment observed scores, like ANCOVAs applied in many RCTs, may result in biased estimates and invalid conclusions. The multigroup model indicated only two items showing a lack of metric and scalar invariance, indicating that between-group comparisons of observed scores are less problematic.

At the same time, according to Meredith (1993), uniqueness invariance is a prerequisite for practical use of tests (e.g., ROM). Our results indicate a substantial lack of uniqueness invariance, in both the longitudinal and multigroup analyses. Therefore, the validity of using BDI scores to estimate or compare treatment effects for individuals should be questioned.

Future Research

The current study is the first to study response shifts in mental health research, using the framework of MI. Although our findings indicate that observed scores cannot be taken as straightforward indicators of treatment efficacy, these findings need replication in different samples. Studying measurement models underlying data sets of other RCTs in this area would provide valuable information on the generalizability of our results. If data of clinical trials consistently show signs of response shift, methodological approaches to counter or take into account response shifts could be implemented in RCTs. For example, providing all participants with the same psychoeducational program may result in comparable response shifts across treatments, and posttreatment measurements in the same metric.

Limitations of the current study could be addressed in future research as well. For example, a sample size of 239 subjects did not allow for estimation of longitudinal models in subgroups. Fitting longitudinal models for treatment groups separately would provide insight into changes in measurement models over the course of different treatments. The question of whether our conclusions, based on data analysis, match the experience of study subjects may be addressed in future research as well. Finally, it should be noted that no invariance at any level could be observed in the current study, according to the widely accepted, more stringent cutoffs of .05 for RMSEA and .95 for CFI (Hu & Bentler, 1999).

In conclusion, our findings indicate that subjects underestimate depressive symptomatology before treatment, compared with after treatment. The item scores of the BDI (and therefore, the BDI total score) become more reliable indicators of depressive symptomatology, and subjects seem to view their symptoms as a more unified concept after treatment. These last two findings may be more evident among subjects receiving psychotherapy, compared with subjects receiving medical treatment.

References

Ahmed, S., Bourbeau, J., Maltais, F., & Mansour, A. (2009). The Oort structural equation modeling approach detected a response shift after a COPD self-management program not detected by the Schmitt technique. *Journal of Clinical Epidemiology, 62,* 1165–1172.

Ahmed, S., Mayo, N., Corbiere, M., Wood-Dauphinee, S., Hanley, J., & Cohen, R. (2005). Change in quality of life of people with stroke over time: True change or response shift? *Quality of Life Research, 14,* 611–627.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 523,* 317–332.

Beck, A. (1979). *Cognitive therapy of depression*. New York, NY: Guilford Press.

Beck, A., & Beamesderfer, A. (1974). Assessment of depression: The

depression inventory. In P. Pichot (Ed.), *Psychological measurements in psychopharmacology: Modern problems in pharmacopsychiatry* (Vol. 7, pp. 151–169). Oxford, England: Karger.

Beck, A., Steer, R., & Garbin, M. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8,* 77–100.

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

Bentler, P. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bollen, K. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Boomsma, A., & Hoogland, J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future. A festschrift in honor of Karl Jöreskog* (pp. 139–168). Lincolnwood, IL: Scientific Software International.

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

Byrne, B. (1995). Strategies in testing for an invariant second-order factor structure: A comparison of EQS and LISREL. *Structural Equation Modeling: A Multidisciplinary Journal, 2,* 53–72.

Byrne, B., & Baron, P. (1993). The Beck Depression Inventory: Testing and cross-validating a hierarchical factor structure for nonclinical adolescents. *Measurement and Evaluation in Counseling and Development, 26,* 164–178.

Byrne, B., & Baron, P. (1994). Measuring adolescent depression: Tests of equivalent factorial structure for English and French versions of the Beck Depression Inventory. *Applied Psychology: An International Review, 43,* 33–47.

Byrne, B., Baron, P., & Balev, J. (1996). The Beck Depression Inventory: Testing for its factorial validity and invariance across gender for Bulgarian non-clinical adolescents. *Personality and Individual Differences, 21,* 641–651.

Byrne, B., Baron, P., & Campbell, T. (1993). Measuring adolescent depression: Factorial validity and invariance of the Beck Depression Inventory across gender. *Journal of Research on Adolescence, 3,* 127–143.

Byrne, B., Baron, P., & Campbell, T. (1994). The Beck Depression Inventory (French version): Testing for gender-invariant factorial structure for nonclinical adolescents. *Journal of Adolescent Research, 9,* 166–179.

Byrne, B., Baron, P., Larsson, B., & Melin, L. (1996). Measuring depression for Swedish nonclinical adolescents: Factorial validity and equivalence of the Beck Depression Inventory across gender. *Scandinavian Journal of Psychology, 37,* 37–45.

Byrne, B., & Campbell, L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology, 30,* 555–574.

Byrne, B., Shavelson, R., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Chen, F., Curran, P., Bollen, K., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36,* 462–494.

Cheung, G., & Rensvold, R. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255.

Cuijpers, P., Geraedts, A., van Oppen, P., Andersson, G., Markowitz, J., & van Straten, A. (2011). Interpersonal psychotherapy for depression: A meta-analysis. *American Journal of Psychiatry, 168,* 581–592.

De Beurs, E., den Hollander-Gijsman, M., Van Rood, Y., van Der Wee, N., Giltay, E., Van Noorden, M., . . . Zitman, F. (2011). Routine outcome monitoring in the Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology & Psychotherapy, 18,* 1–12.

Elkin, I., Parloff, M., Hadley, S., & Autry, J. (1985). NIMH Treatment of Depression Collaborative Research Program. *Archives of General Psychiatry, 42,* 305–316.

Elkin, I., Shea, T., Watkins, J., Imber, S., Sotsky, S., Collins, J., . . . Parloff, M. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. *Archives of General Psychiatry, 46,* 971–982.

Golembiewski, R., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12,* 133–157.

Graham, J. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576.

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

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and itenls. *Applied Psychological Measurement, 9,* 139–164.

Horn, J., & McArdle, J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144.

Hu, L., & Bentler, P. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453.

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

Jöreskog, K. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality & Quantity, 24,* 387–404.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Klerman, G., Weissman, M., Rounsaville, B., & Chevron, E. (1984). *Interpersonal psychotherapy of depression*. Northvale, NJ: Jason Aronson.

Lambert, M., Hansen, N., & Finch, A. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69,* 159–172.

Louks, J., Hayne, C., & Smith, J. (1989). Replicated factor structure of the Beck Depression Inventory. *Journal of Nervous and Mental Disease, 177,* 473–479.

MacCallum, R., Roznowski, M., & Necowitz, L. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

Marsh, H., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum Associates.

Meade, A., Johnson, E., & Braddy, P. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93,* 568–592.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Millsap, R. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R., & Hartog, S. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal of Applied Psychology, 73,* 574–584.

Montgomery, S., & Åsberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry, 134,* 382–389.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Norman, P., & Parker, S. (1996). The interpretation of change in verbal reports: Implications for health psychology. *Psychology & Health, 11,* 301–314.

Oort, F. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14,* 587–598.

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org

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

Revelle, W., & Zinbarg, R. (2009). Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika, 74,* 145–154.

Schafer, J. (2002). norm: Analysis of multivariate normal datasets with missing values (version 1.0–9.2) [Computer software manual]. Retrieved from http://cran.r-project.org/

Schaubroeck, J., & Green, S. (1989). Confirmatory factor analytic procedures for assessing change during organizational entry. *Journal of Applied Psychology, 74,* 892–900.

Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. *Mulitivariate Behavioral Research, 17,* 343–358.

Shafer, A. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology, 62,* 123–146.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120.

Spitzer, R., Endicott, J., & Robins, E. (1978). Research diagnostic criteria: Rationale and reliability. *Archives of General Psychiatry, 35,* 773–782.

Steenkamp, J.-B., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *The Journal of Consumer Research, 25,* 78–90.

Steer, R., Iguchi, M., & Platt, J. (1992). Use of the revised Beck Depression Inventory with intravenous drug users not in treatment. *Psychology of Addictive Behaviors, 6,* 225–232.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50,* 253–263.

Uher, R., Farmer, A., Maier, W., Rietschel, M., Hauser, J., Marusic, A., . . . Aitchison, K. (2008). Measuring depression: Comparison and integration of three scales in the GENDEP study. *Psychological Medicine, 38,* 289–300.

Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70.

Vandenberg, R., & Self, R. (1993). Assessing newcomers' changing commitments to the organization during the first 6 months of work. *Journal of Applied Psychology, 78,* 557–568.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60,* 201–223.

Yoon, M., & Millsap, R. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal, 14,* 435–463.

Zhang, B., Fokkema, M., Cuijpers, P., Li, J., Smits, N., & Beekman, A. (2011). Measurement invariance of the Center for Epidemiological Studies Depression Scale (CES-D) among Chinese and Dutch elderly. *BMC Medical Research Methodology, 11,* 74–85.

Received September 30, 2011
Revision received October 31, 2012
Accepted November 26, 2012 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.