Computerized adaptive testing in Dutch mental health care: a new tool to assess depression and anxiety

Slok-Flens, G.

Chapter 5

Practical Significance of Longitudinal Measurement Invariance Violations in the Dutch-Flemish PROMIS Item Banks for Depression and Anxiety: An Illustration with Ordered-Categorical Data

Published as:

Flens, G., Smits, N., Terwee, C. B., Pijck, L., Spinhoven, P., & de Beurs, E. (2021). Practical Significance of Longitudinal Measurement Invariance Violations in the Dutch–Flemish PROMIS Item Banks for Depression and Anxiety: An Illustration With Ordered-Categorical Data. *Assessment*, *28*(1), 277-294.

5.1 Abstract

We investigated longitudinal measurement invariance in the Dutch–Flemish PROMIS adult v1.0 item banks for Depression and Anxiety using two clinical samples with mood and anxiety disorders (n = 640 and n = 528, respectively). Factor analysis was used to evaluate whether the item banks were sufficiently unidimensional at two test-occasions and whether the measured constructs remained the same over time. The results indicated that the item banks were sufficiently unidimensional, but the thresholds and residual variances of the constructs changed over time. However, using tentative rules of thumb, these invariance violations did not substantially affect the endorsement of a specific response category of a specific item at a specific test-occasion. Furthermore, the impact on the mean latent change scores of the item banks remained below the proposed cutoff value for substantial bias. These findings suggest that the invariance violations lacked practical significance for test users, meaning that the item banks provide sufficiently invariant latent factor scores for use in clinical practice.

Keywords: depression, anxiety, clinical assessment, longitudinal measurement invariance, PROMIS

5.2 Background

In the Netherlands, Dutch-Flemish versions of the Patient-Reported Outcomes Measurement Information System (PROMIS) adult v1.0 item banks for Depression and Anxiety have been developed. In previous studies, the original United States (US) PROMIS adult v1.0 item banks for Depression and Anxiety were translated from English into Dutch-Flemish (Terwee et al., 2014), and psychometrically evaluated for cross-sectional use in both the Dutch general population and ambulatory clinical populations at the start of treatment (Flens et al., 2017, 2019). These studies showed that both item banks have psychometric properties that complied with the PROMIS standards (Reeve et al., 2007). Consequently, adequate item parameters are available that may be used as input for computerized adaptive testing (CAT). CAT is a computer-based method in which items are selected from an item bank based on a respondent's previous item responses. The administration of items stops when a prespecified criterion is met (e.g., a high measurement precision). Consequently, CAT can reduce administration burden with a shorter test while maintaining a high-measurement precision. For more details on CAT, see for example, Embretson and Reise (2000).

Using the Dutch-Flemish PROMIS item banks in CAT simulations, efficient and highly precise measurement of depression and anxiety was obtained (Flens et al., 2017, 2019). Furthermore, the accuracy of the CAT simulations was highly similar compared with that of the full item bank administrations, both in final score estimations and in distinguishing clinical subjects from persons without a mental health disorder. Based on these results, it was concluded that the Dutch-Flemish PROMIS item banks administered by CAT may measure depression and anxiety accurately, precisely, and efficiently in both the general population and clinical ambulatory populations at the start of treatment. When the final goal, however, is to use these CATs in *repeated assessments* of clinical subjects, research also needs to address their longitudinal measurement properties. One of these aspects includes longitudinal measurement invariance (LMI; Widaman, Ferrer, & Conger, 2010).

An item bank is said to be longitudinally measurement invariant when it measures one or more single constructs in the same way over time. This means that changes in test scores of respondents over time can entirely be attributed to changes in the construct(s) measured by the item bank (Fried et al., 2016; Liu et al., 2017). If this is not the case, for example due to the psychoeducation of clinical subjects (Fokkema, Smits, Kelderman, & Cuijpers, 2013; for more explanations, see Fried et al., 2016), then observed changes in test scores are likely to be biased, possibly resulting in wrong inferences about the (change in) construct level. To our knowledge, this kind of bias is investigated in numerous mental health instruments (e.g., Fokkema et al., 2013; Fried et al., 2016; Jabrayilov, Emons, de Jong, & Sijtsma, 2017; te Poel, Hartmann, Baumgartner, & Tanis; 2017), but not yet in any of the PROMIS item banks. The evaluation of LMI in these item banks is highly relevant because in most of the performed LMI studies, it was concluded that the assumption of invariance did not or only partially hold.

In the present study, LMI was investigated for the Dutch-Flemish PROMIS adult v1.0 item banks for Depression and Anxiety using two clinical samples with mood and anxiety disorders respectively. We evaluated whether (a) the item banks were sufficiently unidimensional at two test-occasions, and (b) the measured constructs remained the same over

time. Specifically, LMI was investigated within the framework of factor analysis, using both confirmatory factor analysis (CFA) and exploratory factor analysis (EFA). We modeled the items of the PROMIS item banks explicitly as ordered-categorical. In previous measurement invariance studies, ordered-categorical items were often modeled as continuous because the evaluation of invariance through factor analysis comes with several challenges for ordered-categorial data (Liu et al., 2017; Wu & Estabrook, 2016). Recently, new methodology has become available for CFA which overcomes most of these challenges (Liu et al., 2017). As a result, LMI can be investigated more accurately than would have been the case when the data were modeled as continuous (Rhemtulla, Brosseau-Liard, & Savalei, 2012). In addition, as full LMI rarely holds (van de Schoot et al., 2015), we did not focus solely on statistical significance in the analyses. Additionally, effect sizes based on new methodologies for CFA were evaluated to study the practical significance of the expected invariance violations. Specifically, we investigated two effect sizes that are relevant for test users. This means that we evaluated when (i.e., which test-occasion) and where (i.e., which item and response category) a LMI violation has a substantial impact (Liu et al., 2017), and to what degree changes in test scores are affected (Liu & West, 2018).

5.3 Methods

5.3.1 Participants

Data for this study were collected in two clinical populations that consisted of patients who started ambulant treatment for either a mood disorder or an anxiety disorder. Patients were invited to participate by the Dutch mental health care provider Parnassia Psychiatric Institute, which is the largest mental health institute in the Netherlands and has departments across the country (Flens et al., 2019). Prior to the study, mental health clinicians of Parnassia Psychiatric Institute determined the patient's diagnosis (*DSM-IV*; *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed.; American Psychiatric Association, 1994) with the Dutch translation of the Mini International Neuropsychiatric Interview (i.e., MINI-plus; a structured diagnostic interview used to systematically assess *DSM-IV* diagnoses) in a clinical face-to-face assessment during the intake of treatment. The MINI(-plus) showed sufficient sensitivity, specificity, negative and positive predictive values, and sufficient interrater agreement with other diagnostic instruments; only the interrater agreement on a generalized anxiety disorder and a simple phobia was insufficient (Lecrubier et al., 1997; Muramatsu et al., 2007; Sheehan et al., 1998; van Vliet & de Beurs, 2007). In addition, in accordance with Parnassia Psychiatric Institute's policy, informed consent was obtained before the measurements were administered.

We aimed to include at least 500 patients per sample to be able to adequately examine factor structures (Comrey & Lee, 1992; Liu et al., 2017; MacCallum, Widaman, Zhang, & Hong, 1999). A patient was included when (a) a pretest and posttest were completed without missing item responses, (b) the posttest was administered at least one month after the pretest, and (c) the posttest was administered after the first treatment session. We only included patients that completed a pretest and posttest without missing item responses because our software package (see section Software) could not yet handle missing data using CFA with ordered-

categorical data. For more details on handling missing data in assessing LMI with ordered-categorical data, see Liu et al. (2017). Additionally, the manual of the used software package could be evaluated for any new features (e.g., https://cran.r-project.org/web/packages/lavaan/lavaan.pdf).

5.3.2 Measurements

The measurements consisted of the full Dutch-Flemish PROMIS adult v1.0 item banks for Depression (Flens et al., 2017) and Anxiety (Flens et al., 2019). The Depression item bank was administered to patients who were treated for a mood disorder; the Anxiety item bank was administered to patients whom were treated for an anxiety disorder.

Patients were asked to indicate on a Likert-type scale (1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = *often*, and 5 = *always*) how frequently they experienced a wide range of either depression or anxiety symptoms in the past 7 days. The items reflected symptoms, problems, or negative affective states (e.g., *I felt worthless* for the Depression item bank, or *I felt fearful* for the Anxiety item bank), a higher score meaning more severe depression or anxiety. In Table 5.1 (Depression item bank; 28 items) and Table 5.2 (Anxiety item bank; 29 items), the items with the original PROMIS coding are presented. We should note that the item banks cannot be used without permission of PROMIS (see also www.healthmeasures.net).

For each patient, an item bank was administered digitally through an automated process twice. According to Parnassia Psychiatric Institute's policy, the invitation for the pretest was sent before the intake session. To ensure that at least some treatment was administered and some change in clinical severity would be achieved, the invitation of the posttest was sent at least 1 month after the pretest.

In addition to the administration of the PROMIS item banks, the pretest was preceded by several questionnaires depending on the patient's age and disorder. These questionnaires were not relevant for the purpose of this study and therefore not further described. The posttest administration was not preceded or followed by additional questionnaires.

5.3.3 Statistical analyses

5.3.3.1 Descriptive statistics

The degree of change within patients was evaluated by comparing the mean raw item scores between the pretest and posttest. Uniformity in the pretest to posttest interval was evaluated by calculating quantiles of the days between the pretest and posttest.

5.3.3.2 Unidimensionality

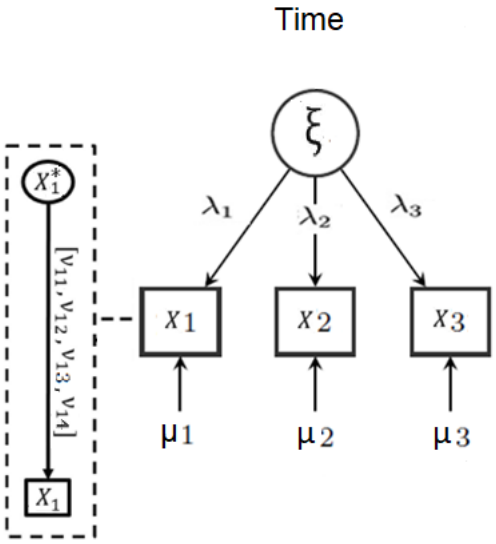
To evaluate LMI in instruments that are theorized to (strongly) reflect one underlying construct, the evaluation of the unidimensionality assumption is a strict condition (Fried et al., 2016). If this assumption is violated, item parameter estimates of CFA will almost inherently be biased, possibly resulting in biased test scores.

Unidimensionality was assessed with EFA as well as CFA (Reeve et al., 2007). With EFA, two factors were extracted from the pretest and posttest data separately. A measurement was considered to be sufficiently unidimensional when the first factor explained more than 20%

of the variance (Reckase, 1979, as cited in Hambleton, 1988), and the ratio of variance explained by the first to second factor was at least 4 (Reeve et al., 2007).

With CFA, a one-factor model was fitted to the pretest and posttest data separately. To illustrate the one-factor CFA model as a first step towards the longitudinal CFA models, it is presented in Figure 5.1, for three example items with five response categories. The model estimates four types of parameters for the ordered-categorial data: (a) *the common factor mean* (ξ) represents the mean of all respondent's latent factor scores; (b) *the factor loadings* (λ) represent for each item the strength and direction of association between the observed item responses and the latent factor scores; (c) *the thresholds* (υ) are cutoff values (the number of thresholds for each item equals the number of response categories minus one) that divide the underlying continuous latent responses into sections, each of which corresponds to endorsing an observed ordinal response category; and (d) *the residual variances* (μ) represent the degree of error with which each item measures the construct of interest. With the resulting model, the degree of unidimensionality was evaluated using the following (scaled [i.e., corrected for nonnormality]) fit-statistics (Fokkema et al., 2013): a scaled CFI ≥ .90 indicates an adequate fit, a scaled CFI ≥ .95 a good fit (Bentler, 1990); a standardized root-mean-square residual (SRMR) ≤ .08 indicates an adequate fit, a SRMR ≤ .05 a good fit (Hu & Bentler, 1999); a scaled root-mean-square error of approximation (RMSEA) ≤ .08 indicates an adequate fit, a scaled RMSEA ≤ .05 a good fit (Browne & Cudeck, 1993).

Figure 5.1 One-factor CFA model for ordered-categorical data with three items and five response categories.

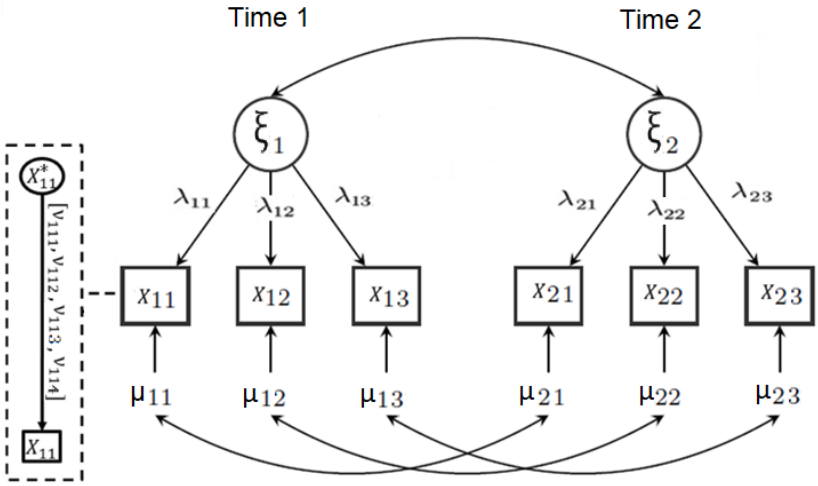


Note. ξ = common factor mean; λ = factor loadings; *X** = continuous latent item responses; *X* = observed item responses; υ = thresholds; μ = residual variances. For each parameter, the first subscript represents the item, the second subscript the threshold number.

5.3.3.3 Tenability of equality constraints

To investigate whether the measured constructs remain the same over time, a series of nested longitudinal CFA models was evaluated and compared (Liu et al., 2017). To illustrate the modeling sequence for evaluating LMI with ordered-categorical data, the general longitudinal model is presented in Figure 5.2, again using three example items and five response categories.

Figure 5.2 Longitudinal CFA model for ordered-categorical data with three items and five response categories.



Note. ξ = common factor mean; λ = factor loadings; *X** = continuous latent item responses; *X* = observed item responses; υ = thresholds; μ = residual variances. For each parameter, the first subscript represents the test-occasion, the second subscript the item number, and the third subscript the threshold number. The longitudinal structure of the model is captured by including a factor correlation between test-occasions as well as a residual correlation between test-occasions for each item.

First, the *baseline invariance model* was fitted. This is a two-factor model in which the pretest and posttest were treated as separate factors. To account for the longitudinal design, a factor correlation was included between test-occasions as well as a residual correlation between test-occasions for each item (Oort, 2005; Vandenberg & Lance, 2000). With the resulting model, it was assessed whether the construct of interest is measured by the same items (i.e., the same content) over time. Second, the baseline invariance model was extended with equality constraints on the factor loadings between test-occasions for each item to create *the loading invariance model*. With this model, it was assessed whether the observed item scores have a similar correlation with the latent factor scores over time. Third, the loading invariance model was extended with equality constraints on the thresholds between test-occasions for each item to create the *threshold invariance model*. With this model, it was assessed whether respondents with similar latent factor scores over time would choose the same response categories. Finally, the threshold invariance model was extended with equality constraints on the residual variances between test-occasions for each item to create the *unique factor invariance model*. With this

model, it was assessed whether the items measure the construct of interest with a similar amount of error over time. Only if this is the case, then an item bank is said to be sufficiently invariant. In other words: equality constraints on factor loadings, thresholds, and residual variances need to be tenable in the longitudinal model to attribute changes in the observed item responses over time entirely to changes in the latent factor over time. A mathematical explanation that supports this can be found in Liu et al. (2017).

To investigate the tenability of the equality constraints, we first evaluated the fit of the longitudinal CFA models using the same fit statistics and cutoff values as for the one-factor CFA models. Second, we compared the fit between two subsequent models with the chi-square (i.e., χ^2) scaled difference test (Satorra, 2000), using an alpha level of .05 to indicate deterioration of fit. Third, because a χ^2 difference test is known to exhibit inflated Type 1 error rates (Sass, Schmitt, & Marsch, 2014), we also evaluated the modification indices of the imposed equality constraints (Liu et al., 2017). When a model showed a modification index above 5, this was considered a deterioration of fit (Jöreskog & Sörbom, 1996). Finally, it has been suggested to also compare the fit between two subsequent models by calculating differences in CFI's or RMSEA's (Chen, Curran, Bollen, Kirby, & Paxton, 2008; Cheung & Rensvold, 2002; Hu & Bentler, 1998). These difference tests, however, have not been properly studied for models with ordered-categorical data (Liu et al., 2017). We therefore chose not to use these fit statistics in our study.

5.3.3.4 CFA model-identification

To be able to estimate the parameters of a CFA model (i.e., to identify the model), some parameters need to be constrained (i.e., the model-identification parameters). For the one-factor CFA models, the following constraints needed to be imposed at each test-occasion (Wu & Estabrook, 2016): (a) the common factor mean was fixed to 0; (b) the factor loading of one single item was fixed to 1; (c) all intercepts were fixed to 0 (intercepts represent the expected item response when the latent factor score is equal to zero, and are only allowed to be estimated when the data is continuous); and (d) all residual variances were fixed to 1. In addition, we needed to impose the following constraints to identify the longitudinal CFA models (Liu et al., 2017): (a) the common factor mean of the pretest was fixed to 0; (b) the factor loading of a single item (i.e., the marker item) was fixed to 1 for both measurements; (c) one threshold of each item and a second threshold for the marker item was constrained to be equal between pretest and posttest; (d) all intercepts were fixed to 0; and (e) all residual variances of the pretest items were fixed to 1.

The constraints on the common factor mean, the intercepts and the residual variances could be imposed directly because *all* parameters were affected within a test-occasion. In the cases of factor loadings and thresholds, however, we needed to impose constraints on *specific* parameters. For these parameters, it is strictly necessary that they are at least longitudinally invariant. Otherwise, baseline invariance will be violated, which will make further model-comparisons biased. In the case of noninvariant threshold model-identification parameters, for example, a true violation of threshold invariance may mistakenly result in the conclusion that loading invariance is violated (Liu et al., 2017). To deal with the possible issue(s) of

noninvariant model-identification parameters, we followed a two-step approach in which the model-identification parameters were selected and evaluated on LMI.

First, the model-identification parameters were selected by comparing the one-factor CFA models of the pretest and posttest based on their factor loading estimates and threshold estimates. Obviously, these models also needed to be identified first before the parameters could be compared. Consequently, we fixed the factor loading of the first item to 1, which is an arbitrary choice. From the remaining items, a marker item was selected based on a trade-off between a high-factor loading estimate for the pretest and posttest (Liu et al., 2017) and a high probability of having an invariant factor loading and two invariant thresholds (i.e., a small difference between the pretest and posttest estimates). Subsequently, we also selected the threshold parameters for the rest of the items based on a high probability of having an invariant threshold (i.e., the thresholds with the smallest difference between the pretest and posttest estimates). The differences between the factor loading estimates as well as the threshold estimates were calculated by subtracting the pretest estimate from the posttest estimate.

Second, we evaluated whether the selected parameters were sufficiently invariant over time. To evaluate LMI, the *baseline invariance model* was compared with the *loading invariance model*. For details about the criteria used to evaluate sufficient invariance, see the section "Tenability of equality constraints over time" above. If these criteria were not met, other parameters were selected for model-identification, and the evaluation of LMI was repeated (Yoon & Millsap, 2007).

5.3.3.5 Practical significance of an invariance violation

When the assumption of LMI is violated, it should be investigated how relevant this violation may be for clinical practice. Specifically, it should be investigated when (i.e., which test-occasion) and where (i.e., which item and response category) the violation has a substantial impact, and to what degree changes in test scores are affected. The findings reveal the usefulness of the measurement to assess change in psychopathology over time. Furthermore, they may help researchers to generate hypotheses as to why the lack of LMI occurs.

5.3.3.5.1 Test-occasions, items, and response categories

Liu and West (2018) proposed to evaluate the practical significance of an invariance violation in ordered-categorical data using two methods. The first methodology is used to investigate to what degree each response category of each item at each measurement occasion is impacted by an invariance violation. To accomplish this, the methodology uses model-predicted probabilities (Liu et al., 2017). These probabilities are *estimations* of the percentage of respondents that endorse each response category on each item at each test-occasion, assuming a specific invariance model. For example, it can be estimated how many respondents would endorse the first item at the pretest with response category *never*, assuming the threshold invariance model. This means that the number of predicted probabilities was 280 for each model of the Depression item bank (2 test-occasions * 28 items * 5 response categories) and 290 for each model of the Anxiety item bank (2 test-occasions * 29 items * 5 response categories).

Model-predicted probabilities were estimated for a model assuming the strictest *achieved* type of LMI (i.e., the strictest model that showed sufficient fit) and a model assuming

a stricter, *violated* type of LMI (i.e., the first model that did not show sufficient fit). We then calculated the differences between the predicted probabilities of the models (i.e., the predicted probability of the model assuming a violated type of LMI minus the predicted probability of the model assuming the strictest achieved type of LMI), which can be considered a measure of the magnitude of an invariance violation. For example, when the model-predicted probability of responding to the first item at the pretest with response category *never* is 20% in the loading invariance model and 30% in the threshold invariance model, 10% of the respondents are predicted to choose a different response category under the assumption of different invariance models. Liu et al. (2017) did not suggest a specific cutoff value to interpret this difference, but they used a difference of 5% as illustration for a small impact based on 749 respondents. We chose to follow their example, meaning that when this cutoff was exceeded, more than 5% of the patients are expected to choose a different response category for a specific item at a specific test-occasion, assuming different types of LMI.

5.3.3.5.2 Mean latent change score

The second methodology to evaluate the practical significance of an invariance violation in ordered-categorical data uses the *estimated* mean latent change score (i.e., the difference between the estimated common factor means of the pretest and posttest). This methodology was introduced by Liu and West (2018) for a specific type of longitudinal model (i.e., the latent growth model) and can be extended to the longitudinal model with two factors. This application consists of three steps.

The first step was to create a subset of items (i.e., the anchor set) that showed a specific type of LMI sufficiently (e.g., threshold invariance), which would be used in the second step to evaluate the remaining items on that type of LMI. Some authors use all items (or all items except one) to evaluate them individually on a specific type of LMI, but it has been shown to be more accurate if these evaluations are solely based on a group of invariant items (i.e., between 10% and 20% of the full item set; Woods, 2009). We therefore created an anchor set using the following steps. First, all items were evaluated individually with the χ^2 scaled difference test by comparing the model assuming a stricter, violated type of LMI to the same model minus the model-specific equality constraint(s) for 1 item. Next, 20% of the items (i.e., six items for both item banks) were selected as anchor set based on the lowest $\Delta\chi^2$ values. Finally, it was evaluated with the χ^2 scaled difference test whether the anchor set was sufficiently invariant. This was done by comparing the model assuming the strictest achieved type of LMI (e.g., the loading invariance model) to the same model including equality constraints on the anchor set (in our example that would be the inclusion of equality constraints on the thresholds of the anchor set items). If the anchor set did not show sufficient invariance, we removed the additional equality constraint(s) from the item that earlier showed the highest $\Delta\chi^2$ statistic of the anchor set items, and repeated the evaluation of LMI.

The second step was to establish which additional items showed a specific type of LMI sufficiently. To accomplish this, all items were evaluated individually with the χ^2 scaled difference test by comparing the model assuming the strictest achieved type of LMI including equality constraints on the anchor set with the same model including the equality constraint(s) on one additional item. For example, to evaluate which additional items showed sufficient

threshold invariance, the loading invariance model including threshold constraints on the anchor set was compared with the same model including threshold constraints on one additional item, and this was repeated for all items.

The third and final step was to assess the impact of an invariance violation on the mean latent change score. To accomplish this, the *relative mean change* was calculated between the model with equality constraints on all invariant items (i.e., the *partial invariance model*) and the model assuming a stricter, violated type of LMI (i.e., the *full invariance model*). This relative mean change was calculated as the difference between the mean latent change score of these two models, divided by the mean latent change score of the full invariance model. As mean latent change score, we used the estimated common factor mean of the posttest as this equals the mean latent change score in a longitudinal model in which the common factor mean of the pretest was set to 0 for model-identification purposes. Following the suggestion of Kaplan (1989, as cited in Flora & Curran, 2004), a relative mean change value larger than 10% was considered as indicative of substantial bias. When this was the case, the modeling sequence was continued with the partial invariance model. Otherwise, the modeling sequence was continued with the full invariance model.

5.3.3.6 Software

We performed all analyses separately for the Depression and the Anxiety item banks in the statistical environment R (R Core Team, 2017). EFA was conducted with the R package psych (Version 1.5.4; Revelle, 2013); CFA was conducted with the R package lavaan (Version 0.5-18; Rosseel, 2012) using theta parametrization and the diagonally weighted least squares estimator with robust standard errors and a mean and variance adjustment (i.e., WLSMV in lavaan; Liu et al., 2017). Furthermore, both factor analyses were conducted on the polychoric correlation matrix (Bollen, 1989). For some guidelines on selecting a software package, an estimation method, and a framework for analysis (i.e., factor analysis or item response theory [IRT]) for investigating LMI with ordered-categorical data, see Li, 2016; Liu et al., 2017).

5.4 Results

5.4.1 Demographic characteristics

The eligible sample consisted of 13,802 patients (Depression, *n* = 8,372; Anxiety, *n* = 5,430). Of these patients, 13,067 (Depression, *n* = 7,715; Anxiety, *n* = 5,352) were invited to respond to the pretest and 5,383 (Depression, *n* = 3,031; Anxiety, *n* = 2,352) also completed it (pretest response rate Depression item bank = 39.3%; pretest response rate Anxiety item bank = 43.9%). Of the patients with completed pretests, 2,962 patients (Depression, *n* = 1,561; Anxiety, *n* = 1,401) were invited to respond to the posttest and 1,253 patients (Depression, *n* = 664; Anxiety, *n* = 589) also completed it (posttest response rate Depression item bank = 42.5%; posttest response rate Anxiety item bank = 42.0%). None of the patients with a completed pretest and posttest had specific missing item responses. Consequently, we did not have to exclude any more patients for not meeting our first inclusion criterium. We did exclude 85 more patients for not meeting the remaining inclusion criteria (i.e., the posttest was administered less than one

month after the pretest, and/or before the first treatment session). Our final study sample therefore consisted of n = 640 for the Depression item bank (total sample response rate = 7.6%; 62% female; M_{age} = 41.3 years, SD = 13.4, range 18–77) and n = 528 for the Anxiety item bank (total sample response rate = 9.7%; 65% female; M_{age} = 37.1 years, SD = 12.9, range 18–73). These final samples did not contain sparse data (i.e., missing specific item response *categories* within items). Consequently, LMI could be investigated in a straightforward fashion (Liu et al., 2017).

Table 5.1 Item *M* (*SD*) for the pretest and posttest of the Depression item bank.

Item code	Item	Pretest	Posttest
EDDEP04	I felt worthless	3.44 (1.04)	2.88 (1.06)
EDDEP05	I felt that I had nothing to look forward to	3.48 (1.09)	2.95 (1.16)
EDDEP06	I felt helpless	3.42 (1.01)	2.93 (1.09)
EDDEP07	I withdrew from other people	3.57 (0.93)	3.10 (1.05)
EDDEP09	I felt that nothing could cheer me up	3.49 (0.96)	2.97 (1.11)
EDDEP14	I felt that I was not as good as other people	3.53 (1.10)	3.03 (1.14)
EDDEP17	I felt sad	3.75 (0.89)	3.25 (1.04)
EDDEP19	I felt that I wanted to give up on everything	3.32 (1.05)	2.80 (1.12)
EDDEP21	I felt that I was to blame for things	3.23 (1.11)	2.78 (1.12)
EDDEP22	I felt like a failure	3.28 (1.19)	2.82 (1.18)
EDDEP23	I had trouble feeling close to people	3.20 (1.12)	2.89 (1.14)
EDDEP26	I felt disappointed in myself	3.68 (1.04)	3.20 (1.12)
EDDEP27	I felt that I was not needed	3.35 (1.14)	2.95 (1.18)
EDDEP28	I felt lonely	3.64 (1.08)	3.18 (1.18)
EDDEP29	I felt depressed	3.85 (1.02)	3.15 (1.20)
EDDEP30	I had trouble making decisions	3.57 (0.98)	3.08 (1.12)
EDDEP31	I felt discouraged about the future	3.76 (1.05)	3.20 (1.23)
EDDEP35	I found that things in my life were overwhelming	3.28 (1.11)	2.88 (1.14)
EDDEP36	I felt unhappy	3.78 (1.00)	3.18 (1.13)
EDDEP39	I felt I had no reason for living	2.72 (1.31)	2.28 (1.23)
EDDEP41	I felt hopeless	3.19 (1.09)	2.75 (1.16)
EDDEP42	I felt ignored by people	2.80 (1.05)	2.52 (1.06)
EDDEP44	I felt upset for no reason	3.12 (1.08)	2.70 (1.09)
EDDEP45	I felt that nothing was interesting	3.37 (1.04)	2.87 (1.13)
EDDEP46	I felt pessimistic	3.44 (1.02)	2.98 (1.12)
EDDEP48	I felt that my life was empty	3.43 (1.13)	2.91 (1.23)
EDDEP50	I felt guilty	3.36 (1.13)	2.90 (1.16)
EDDEP54	I felt emotionally exhausted	3.85 (1.06)	3.28 (1.21)

As the response rates were small, additional tests were performed for each item bank to examine whether the composition of the included patients was similar to that of the nonincluded

patients. For the variable gender, we investigated the effect size Pearson's residual, following the suggestion of 2.00 as cutoff value for indicating a systematic difference between the observed and expected number of respondents (Agresti & Kateri, 2011). For the variables age and pretest score (i.e., the sum of the item scores), we investigated the effect size Cohen's *d* (i.e., the difference between the mean ages/pretest scores divided by the pooled *SD*), following the guideline proposed by Cohen to interpret the size of the effect (1988): 0.20 = small effect, 0.50 = medium effect, 0.80 = large effect. The results showed for both item banks that Pearson's residuals were all below 2.00 and Cohen's *d*s were below 0.20. We therefore concluded that the included patients for each item bank did not differ substantially from the nonincluded patients regarding the variables gender, age and pretest score.

Table 5.2 Item *M* (*SD*) for the pretest and posttest of the Anxiety item bank.

Item code	Item	Pretest	Posttest
EDANX01	I felt fearful	3.57 (0.86)	3.10 (0.96)
EDANX02	I felt frightened	2.92 (1.11)	2.47 (1.06)
EDANX03	It scared me when I felt nervous	3.16 (1.15)	2.89 (1.04)
EDANX05	I felt anxious	3.54 (0.94)	3.11 (0.96)
EDANX07	I felt like I needed help for my anxiety	3.51 (1.12)	2.81 (1.12)
EDANX08	I was concerned about my mental health	3.29 (1.15)	2.78 (1.12)
EDANX12	I felt upset	3.21 (1.05)	2.83 (1.05)
EDANX13	I had a racing or pounding heart	2.90 (1.16)	2.61 (1.11)
EDANX16	I was anxious if my normal routine was disturbed	2.92 (1.22)	2.65 (1.17)
EDANX18	I had sudden feelings of panic	3.06 (1.19)	2.59 (1.12)
EDANX20	I was easily startled	2.71 (1.23)	2.39 (1.13)
EDANX21	I had trouble paying attention	3.07 (1.10)	2.88 (1.12)
EDANX24	I avoided public places or activities	2.63 (1.33)	2.34 (1.24)
EDANX26	I felt fidgety	3.74 (0.97)	3.32 (1.05)
EDANX27	I felt something awful would happen	2.64 (1.24)	2.29 (1.15)
EDANX30	I felt worried	3.72 (0.95)	3.26 (1.02)
EDANX33	I felt terrified	2.36 (1.23)	1.98 (1.07)
EDANX37	I worried about other people's reactions to me	3.13 (1.23)	2.82 (1.22)
EDANX40	I found it hard to focus on anything other than my anxiety	3.25 (1.13)	2.83 (1.14)
EDANX41	My worries overwhelmed me	3.01 (1.19)	2.53 (1.21)
EDANX44	I had twitching or trembling muscles	2.32 (1.17)	2.15 (1.08)
EDANX46	I felt nervous	3.47 (0.96)	3.14 (0.96)
EDANX47	I felt indecisive	3.15 (1.12)	2.80 (1.11)
EDANX48	Many situations made me worry	3.22 (1.06)	2.81 (1.11)
EDANX49	I had difficulty sleeping	3.22 (1.31)	2.91 (1.29)
EDANX51	I had trouble relaxing	3.73 (0.98)	3.31 (1.12)
EDANX53	I felt uneasy	3.28 (1.00)	2.96 (1.07)
EDANX54	I felt tense	3.74 (0.91)	3.36 (1.00)
EDANX55	I had difficulty calming down	3.15 (1.06)	2.77 (1.12)

5.4.2 Descriptive statistics

Table 5.1 (Depression item bank) and Table 5.2 (Anxiety item bank) display the mean item scores (*SD*'s) of the pretest and posttest. All items showed a decrease in mean from pretest to posttest, ranging from 0.27 to 0.71 for the Depression item bank and from 0.17 to 0.70 for the Anxiety item bank.

Concerning the pretest to posttest interval, the median was 238.50 days for the Depression item bank (range = 43.00–803.00, interquartile range = 219.00–281.00) and 181.50 days for the Anxiety item bank (range = 39.00 – 825.00, interquartile range = 158.00–278.25). These results indicate that the degree of uniformity in the pretest to posttest interval was quite low for both item banks.

5.4.3 Model-identification parameters

For the *Depression item bank*, item EDDEP05 (i.e., *I felt that I had nothing to look forward to*) was selected as marker item because it showed a large factor loading for both pretest ($\lambda_2 = 0.93$) and posttest ($\lambda_2 = 0.93$) that did not differ between test-occasions. Furthermore, we found relatively moderate differences between the test-occasions in the first and second threshold of this item ($\Delta\upsilon_1 = 0.41$, $\Delta\upsilon_2 = 0.78$). In addition, we selected the first threshold of the remaining items for showing the smallest difference between the test-occasions' estimates.

The evaluation of LMI in the selected parameters showed that the loading invariance model was not rejected by the χ^2 scaled difference test (see Table 5.3, line 1 and line 2 of the Depression item bank). Furthermore, all modification indices of the constrained parameters were below 5. We concluded that the selected parameters of the Depression item bank were sufficiently invariant for model-identification.

Table 5.3 Fit statistics for the longitudinal CFA (invariance) models of the Depression and Anxiety item banks.

Item bank	Invariance model	*df*	χ^2	Δdf	$\Delta\chi^2$	*p*	CFI	SRMR	RMSEA
Depression	Baseline	1455	5449.131	-	-	-	0.955	0.051	0.057
	Loading	1482	5472.148	21.165	30.087	0.094	0.955	0.051	0.056
	Threshold	1565	5612.895	52.499	144.376	0.000	0.954	0.051	0.055
	Unique Factor	1593	6068.212	22.635	94.926	0.000	0.956	0.052	0.053
Anxiety	Baseline	1565	5006.248	-	-	-	0.954	0.055	0.054
	Loading	1593	5035.036	22.580	33.380	0.067	0.954	0.055	0.053
	Threshold	1679	5206.926	50.578	144.475	0.000	0.953	0.055	0.052
	Factor Variance	1708	5656.931	23.699	89.184	0.000	0.955	0.057	0.051

Note. *df* = degrees of freedom; χ^2 = unscaled chi-square; Δdf = scaled difference in degrees of freedom based on the preceding model; $\Delta\chi^2$ = scaled difference in chi-square based on the preceding model; *p* = *p*-value for the chi-square scaled difference test; CFI = scaled comparative fit index; SRMR = standardized root-mean-square residual; RMSEA = scaled root-mean-square error of approximation.

For the *Anxiety item bank*, we selected item EDANX40 (i.e., *I found it hard to focus on anything other than my anxiety*) as marker item because the factor loading was adequate for the pretest ($\lambda_2 = 0.69$) and posttest ($\lambda_2 = 0.77$) and differed only somewhat between test-occasions ($\Delta\lambda_2 = 0.08$). Furthermore, we found relatively moderate differences between the test-occasions in the first and second threshold ($\Delta\upsilon 1 = 0.34$, $\Delta\upsilon 2 = 0.57$). In addition, we selected the first threshold for almost all remaining items because the difference between the test-occasions' estimates was the smallest, except for items EDANX03, EDANX21, and EDANX46, for which the smallest difference was found for the second threshold.

The evaluation of LMI in the selected parameters showed that the loading invariance model was rejected. Furthermore, the modification indices of the constrained parameters were above 5 for both the factor loading and the first threshold of item EDANX05 (i.e., *I felt anxious*). When we changed the equality constraint of this item from the first to the second threshold, the loading invariance model was no longer rejected (see Table 5.3, line 1 and line 2 of the Anxiety item bank). Moreover, the modification indices of the constrained parameters were all below 5. We concluded that the (adjusted) selection of parameters for the Anxiety item bank were sufficiently invariant for model-identification.

5.4.4 Unidimensionality of the item banks

EFA showed that the first and second factor of the pretest explained 58% and 6% of the variance for the Depression item bank, and 54% and 6% for the Anxiety item bank, respectively. For the posttest, the first and second factor explained 68% and 4% of the variance for the Depression item bank, and 63% and 5% for the Anxiety item bank, respectively. The variances explained by the first factor were above 20% and the ratios of variance explained by the first to second factor were larger than 4. Both item banks were therefore considered to be sufficiently unidimensional at both measurements. Moreover, as both indices of unidimensionality improved from pretest to posttest, the constructs Depression and Anxiety can be considered to become more homogeneous over time.

Table 5.4 Fit statistics for the one-factor CFA models of the Depression and Anxiety item banks.

Item bank	Measurement	*df*	CFI	SRMR	RMSEA
Depression	Pretest	350	0.916	0.063	0.111
	Posttest	350	0.964	0.042	0.097
Anxiety	Pretest	377	0.910	0.067	0.106
	Posttest	377	0.959	0.052	0.094

Note. *df* = degrees of freedom; CFI = scaled comparative fit index; SRMR = standardized root-mean-square residual; RMSEA = scaled root-mean-square error of approximation.

In Table 5.4, the fit statistics are presented for all evaluated one-factor CFA models. For the pretest, the CFI and SRMR indicated adequate model fit for both item banks; the RMSEA indicated a moderate fit. For the posttest, the model fit improved for both item banks according

to all fit statistics. Moreover, the fit changed from adequate to good for the CFI of both item banks and the SRMR of the Depression item bank. These results are in line with the findings of EFA: the item banks showed sufficient unidimensionality at both test-occasions, and the constructs Depression and Anxiety became more homogeneous over time.

5.4.5 Tenability of equality constraints

In Table 5.3, the fit statistics are presented for all evaluated longitudinal CFA models. The results were highly similar for both item banks. According to the CFI, SRMR, and RMSEA, all models showed good model fit. The χ^2 scaled difference test showed that including constraints on factor loadings did not worsen the model fit, but including constraints on thresholds and residual variances did worsen the model fit. Furthermore, for the Depression item bank, modification indices above 5 were found for threshold constraints of 8 items and residual variance constraints of 10 items. For the Anxiety item bank, modification indices above 5 were found for threshold constraints of 9 items and residual variance constraints of 10 items. These results indicate that equality constraints on factor loadings were tenable in the longitudinal model, but equality constraints on thresholds and residual variances were not tenable. In other words, we found for both item banks that loading invariance was achieved, but threshold invariance and unique factor invariance were violated.

5.4.6 The magnitude and practical significance of the invariance violations

5.4.6.1 Threshold invariance

In Table 5.5 (Depression item bank) and Table 5.6 (Anxiety item bank), the differences are presented between the model-predicted probabilities of the loading invariance model and the threshold invariance model. For the Depression item bank, all of the 280 differences were below the cutoff value of 5%. Both the lowest and highest difference were found for response Category 4 (i.e., *often*) of item EDDEP17 (*I felt sad*). The number of respondents that are predicted to endorse this response category on this item at the pretest was 3.9% lower in the threshold model than in the loading invariance model, while at the posttest it was 3.8% higher. In addition, for the Anxiety item bank, only 2 out of 290 differences were somewhat above the cutoff value of 5%. The number of respondents that are predicted to endorse response Category 2 (i.e., *rarely*) on item EDANX07 (i.e., *I felt like I needed help for my anxiety*) at the pretest was 6.1% higher in the threshold model than in the loading invariance model while it was 5.6% lower at the posttest. Consequently, the *overall* results indicate that the rejection of threshold invariance does not substantially affect the endorsement of a specific response category of a specific item administered at a specific test-occasion.

To evaluate to what extent the mean latent change score was impacted by the threshold invariance violation, an anchor set was first created for each item bank. We selected items EDDEP05, EDDEP21, EDDEP28, EDDEP31, EDDEP35, and EDDEP48 as anchor set for the Depression item bank, and items EDANX12, EDANX20, EDANX40, EDANX41, EDANX46, and EDANX49 for the Anxiety item bank. Both of these item sets showed sufficient threshold invariance according to the χ^2 scaled difference test. When we used these item sets to evaluate the other items on threshold invariance, items EDDEP04, EDDEP06, EDDEP07, EDDEP09, EDDEP17, EDDEP23, EDDEP29, EDDEP30, EDDEP36, EDDEP46, and EDDEP54 did not

Table 5.5 Differences between the model-predicted probabilities of choosing specific response categories on specific items at specific test-occasions based on the loading invariance and the threshold models for the Depression item bank.

	Never		Rarely		Sometimes		Often		Always	
Item code	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
EDDEP04	-0,005	0,006	0,009	-0,010	0,016	-0,016	-0,006	0,006	-0,014	0,014
EDDEP05	-0,005	0,006	-0,011	0,013	0,027	-0,031	-0,008	0,009	-0,002	0,002
EDDEP06	-0,005	0,006	-0,002	0,002	0,030	-0,031	-0,036	0,034	0,013	-0,011
EDDEP07	-0,002	0,003	0,015	-0,018	0,014	-0,014	-0,029	0,030	0,002	-0,002
EDDEP09	-0,005	0,005	0,027	-0,030	-0,018	0,020	-0,017	0,015	0,013	-0,010
EDDEP14	-0,004	0,005	0,003	-0,004	0,016	-0,016	-0,011	0,011	-0,004	0,004
EDDEP17	-0,002	0,002	0,018	-0,019	0,017	-0,016	-0,039	0,038	0,005	-0,005
EDDEP19	-0,006	0,007	0,017	-0,020	0,004	-0,003	-0,019	0,019	0,003	-0,003
EDDEP21	-0,005	0,006	0,000	-0,001	0,008	-0,009	-0,004	0,004	0,001	-0,001
EDDEP22	-0,008	0,010	-0,018	0,020	0,029	-0,033	-0,003	0,003	0,000	0,000
EDDEP23	-0,005	0,006	-0,010	0,010	0,001	-0,003	-0,010	0,005	0,023	-0,018
EDDEP26	-0,003	0,003	-0,008	0,009	0,013	-0,015	0,002	-0,002	-0,005	0,004
EDDEP27	-0,006	0,006	-0,001	0,001	-0,011	0,010	0,007	-0,008	0,011	-0,009
EDDEP28	-0,003	0,004	0,006	-0,007	-0,002	0,002	-0,011	0,010	0,010	-0,009
EDDEP29	-0,004	0,005	0,020	-0,024	0,028	-0,029	-0,029	0,033	-0,014	0,014
EDDEP30	-0,002	0,003	0,030	-0,032	-0,020	0,022	-0,012	0,012	0,005	-0,004
EDDEP31	-0,004	0,005	0,008	-0,010	0,001	-0,001	-0,006	0,007	0,001	-0,001
EDDEP35	-0,005	0,005	0,006	-0,007	-0,018	0,016	0,013	-0,012	0,003	-0,003
EDDEP36	-0,003	0,004	0,000	0,000	0,022	-0,025	-0,007	0,009	-0,012	0,013
EDDEP39	-0,019	0,021	0,010	-0,011	0,004	-0,005	-0,001	0,001	0,007	-0,006
EDDEP41	-0,009	0,011	-0,008	0,009	0,003	-0,005	0,010	-0,011	0,004	-0,004
EDDEP42	-0,007	0,008	0,008	-0,008	-0,022	0,020	0,012	-0,012	0,010	-0,007
EDDEP44	-0,006	0,006	0,024	-0,024	-0,014	0,015	-0,007	0,006	0,003	-0,003
EDDEP45	-0,005	0,006	0,017	-0,019	-0,014	0,015	-0,004	0,004	0,006	-0,005
EDDEP46	-0,004	0,005	0,011	-0,012	0,011	-0,010	-0,033	0,031	0,016	-0,013
EDDEP48	-0,007	0,010	0,003	-0,005	0,004	-0,005	-0,001	0,001	0,002	-0,002
EDDEP50	-0,004	0,005	-0,009	0,012	0,031	-0,037	-0,007	0,008	-0,010	0,012
EDDEP54	-0,004	0,004	0,010	-0,011	0,031	-0,032	-0,031	0,033	-0,007	0,006

Note. T1 = pretest; T2 = posttest; each difference is based on the model-predicted probability of the threshold invariance model minus the model-predicted probability of the loading invariance model.

show sufficient invariance for the Depression item bank, and items EDANX01, EDANX03, EDANX05, EDANX07, EDANX08, EDANX18, EDANX26, EDANX30, EDANX51, and EDANX53 did not show sufficient invariance for the Anxiety item bank. However, the relative mean change between the full threshold invariance model and the partial threshold invariance model did not exceed the cutoff value of 10% for both item banks (although that of the Anxiety item bank came close to 10%). For the Depression item bank, the mean latent change score was -0.81 for the full threshold invariance model and -0.76 for the partial threshold invariance model, resulting in a relative mean change of 6.82%. For the Anxiety item bank, the mean latent

change score was -0.61 for the full threshold invariance model and -0.55 for the partial threshold invariance model, resulting in a relative mean change of 9.58%. These results indicate that the bias caused by the threshold invariance violation on the mean latent change score was not substantial for both item banks. Consequently, we decided to continue the modeling sequence for both item banks using the full threshold invariance model.

Table 5.6 Differences between the model-predicted probabilities of choosing specific response categories on specific items at specific test-occasions based on the loading invariance and the threshold models for the Anxiety item bank.

	Never		Rarely		Sometimes		Often		Always	
Item code	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
EDANX01	0.001	-0.002	0.015	-0.015	0.018	-0.017	-0.030	0.032	-0.004	0.004
EDANX02	-0.007	0.008	0.021	-0.022	-0.018	0.018	0.013	-0.013	-0.008	0.009
EDANX03	-0.032	0.038	0.014	-0.020	0.011	-0.011	0.007	-0.007	-0.001	0.001
EDANX05	-0.016	0.022	0.009	-0.012	0.037	-0.041	-0.024	0.025	-0.006	0.006
EDANX07	0.000	0.002	0.061	-0.056	0.010	-0.010	-0.050	0.045	-0.021	0.019
EDANX08	0.000	0.002	0.018	-0.019	0.020	-0.017	-0.034	0.031	-0.003	0.002
EDANX12	-0.001	0.000	0.006	-0.005	-0.001	0.002	-0.002	0.002	-0.002	0.002
EDANX13	-0.003	0.002	-0.008	0.009	0.002	-0.002	0.013	-0.012	-0.004	0.003
EDANX16	-0.004	0.003	-0.017	0.017	0.010	-0.010	0.003	-0.003	0.008	-0.006
EDANX18	-0.007	0.007	0.011	-0.011	0.026	-0.026	-0.023	0.023	-0.007	0.007
EDANX20	-0.009	0.008	-0.003	0.003	0.012	-0.011	0.003	-0.003	-0.003	0.003
EDANX21	-0.013	0.015	-0.004	0.001	-0.005	0.005	0.011	-0.012	0.010	-0.009
EDANX24	-0.016	0.015	0.006	-0.005	0.003	-0.003	0.001	-0.001	0.007	-0.006
EDANX26	0.001	-0.004	0.014	-0.012	-0.008	0.008	0.000	0.001	-0.007	0.007
EDANX27	-0.016	0.016	0.000	-0.001	0.013	-0.012	0.000	0.000	0.003	-0.003
EDANX30	0.002	-0.003	0.012	-0.012	0.021	-0.021	-0.025	0.025	-0.011	0.010
EDANX33	-0.029	0.031	0.020	-0.023	0.006	-0.006	0.010	-0.010	-0.007	0.008
EDANX37	0.003	-0.001	-0.009	0.007	0.015	-0.014	-0.018	0.015	0.010	-0.008
EDANX40	-0.003	0.002	-0.015	0.018	0.022	-0.025	-0.005	0.004	0.001	-0.001
EDANX41	-0.009	0.011	0.016	-0.019	-0.007	0.007	-0.006	0.005	0.006	-0.005
EDANX44	-0.019	0.016	-0.003	0.004	0.009	-0.010	0.013	-0.012	-0.001	0.001
EDANX46	-0.005	0.005	-0.003	0.003	0.008	-0.009	0.000	0.000	0.000	0.000
EDANX47	0.000	0.000	0.003	-0.003	0.008	-0.007	-0.023	0.019	0.012	-0.009
EDANX48	-0.001	0.003	0.009	-0.010	-0.009	0.009	-0.007	0.005	0.008	-0.006
EDANX49	0.000	0.000	0.004	-0.004	-0.013	0.012	0.001	-0.002	0.009	-0.007
EDANX51	0.002	-0.001	0.026	-0.026	-0.028	0.026	-0.009	0.008	0.009	-0.007
EDANX53	0.001	0.001	-0.002	0.001	-0.006	0.005	-0.019	0.011	0.027	-0.018
EDANX54	0.001	-0.003	0.015	-0.014	-0.020	0.020	-0.003	0.002	0.006	-0.005
EDANX55	-0.003	0.004	0.009	-0.009	-0.015	0.014	-0.005	0.002	0.014	-0.010

Note. T1 = pretest; T2 = posttest; each difference is based on the model-predicted probability of the threshold invariance model minus the model-predicted probability of the loading invariance model.

5.4.6.2 Unique factor invariance

For the unique factor invariance violation, all differences between the model-predicted probabilities of the threshold invariance model and the unique factor invariance model did not exceed the cutoff value of 5% for both item banks. The differences were found to be between -2.7% and 2.6% for the Depression item bank, and between -3.7% and 3.2% for the Anxiety item bank. Consequently, the overall results indicate that the rejection of unique factor invariance does not substantially affect the endorsement of a specific response category of a specific item administered at a specific test-occasion.

In addition, we selected items EDDEP19, EDDEP29, EDDEP30, EDDEP41, EDDEP42, and EDDEP54 as anchor set for the Depression item bank, and items EDANX12, EDANX13, EDANX24, EDANX26, EDANX37, and EDANX41 as anchor set for the Anxiety item bank. The item set of the Depression item bank showed sufficient invariance according to the χ^2 scaled difference test. For the Anxiety item bank, however, we had to remove the equality constraints of item EDANX37 and EDANX41 before the anchor item set was sufficiently invariant. When we used these item sets to evaluate the other items on unique factor invariance, items EDDEP04, EDDEP06, EDDEP09, EDDEP17, EDDEP23, EDDEP27, EDDEP28, EDDEP35, EDDEP44, EDDEP45, and EDDEP50 did not show sufficient invariance for the Depression item bank, and items EDANX03, EDANX07, EDANX08, EDANX27, EDANX46, EDANX47, EDANX48, EDANX51, EDANX53, EDANX54, and EDANX55 did not show sufficient invariance for the Anxiety item bank. However, the relative mean change between the full unique factor invariance model and the partial unique factor invariance model did not exceed the cutoff value of 10% for both item banks. For the Depression item bank, the mean latent change score was -0.84 for the full unique factor invariance model and -0.85 for the partial unique factor invariance model, resulting in a relative mean change of -1.88%. For the Anxiety item bank, the mean latent change score was -0.65 for the full unique factor invariance model and -0.64 for the partial unique factor invariance model, resulting in a relative mean change of 2.04%. These results indicate that the bias caused by the unique factor invariance violation was not substantial for the mean latent change score of both item banks.

5.5 Discussion

Until now, none of the PROMIS item banks were evaluated on LMI. In the present study, LMI was investigated in the Dutch-Flemish PROMIS adult v1.0 item banks for Depression and Anxiety using two clinical samples with mood and anxiety disorders. To study LMI, we used factor analysis to evaluate whether (a) the item banks were sufficiently unidimensional at two test-occasions, and (b) the measured constructs remained the same over time. Moreover, we assessed two effect sizes relevant for test users to evaluate the practical significance of the found invariance violations. Specifically, we investigated when (i.e., which test-occasion) and where (i.e., which item and response category) the LMI violations had a substantial impact (Liu et al., 2017), and to what degree changes in test scores were affected (Liu & West, 2018).

Both EFA and one-factor CFA indicated that the item banks were sufficiently unidimensional. The measured constructs, however, became more homogeneous over time, indicating some change within the constructs. Longitudinal CFA models confirmed this change in the constructs as equality constraints on thresholds and residual variances were shown to be untenable. These results indicate that the item banks may lead to biased pretest to posttest change scores. Similar results were found by Fokkema et al. (2013) and Fried et al. (2016) for other instruments measuring depression.

We performed two analyses to gauge the practical significance of the invariance violations using tentatively determined rules of thumb. In the first analysis, we found that none of the response categories of each item at each test-occasion was substantially affected by the violations. Only the Anxiety item bank showed that the number of respondents predicted to endorse response Category 2 (i.e., *rarely*) on item EDANX07 (i.e., *I felt like I needed help for my anxiety*) at the pretest was 6.1% higher in the threshold invariance model than in the loading invariance model, while at the posttest it was 5.6% lower. This item is included in two out of four of the PROMIS short-forms (i.e., short-form 6a and 8a), but because the differences can be considered somewhat small, the impact on scores will likely be small. In addition, the second practical significance analysis showed that none of the relative mean changes between the estimated mean latent change scores of the pretest and posttest exceeded our cutoff value for substantial bias. These results suggest that the item banks provide sufficiently invariant latent factor scores for use in clinical practice. We should stress, however, that the practical significance analysis of Liu et al. (2017) still needs to be investigated further to confirm that it is equally sensitive to invariance violations of factor loadings, thresholds, and residual variances. Moreover, the detection of individual (non)invariant items, performed in the practical significance analysis of Liu and West (2018), is complex and many procedures are, to some extent, conceptually or statistically flawed (Bechger & Maris, 2015; Borsboom, 2016). Therefore, we cannot rule out that the Dutch-Flemish PROMIS item banks for Depression and Anxiety lack LMI to at least some degree for patients with a mood and anxiety disorder. In particular, the Anxiety item bank may be vulnerable for LMI, as the relative mean change for the threshold invariance violation came close to the proposed cutoff value for substantial bias. Thus, the mean latent change score may not entirely represent actual changes in the constructs over time as measured through the item banks.

Assuming at least some invariance violations, Fried et al. (2016) argued that possible problems with LMI do not imply that test scores are not useful in clinical practice or that they should not be interpreted, as we can safely assume that the sum of symptoms does provide information about the general psychopathological burden people carry. This means that when an instrument shows practically significant invariance violations, it may still be used to assess clinical subjects meaningfully, albeit with somewhat more caution. Furthermore, in the case of assessing individuals, a test user should be aware that an instrument is a tool designed to help practitioners as a complement to their clinical expertise and not as an objective decision tool (i.e., each test-score includes measurement error; Greenhalgh et al., 2018). Therefore, professionals should not only discuss (changes in) test scores with their patients, but also question them on the development of specific symptoms and the progress towards their treatment goals. In addition, when assessing groups, researchers should decide whether the

possible bias due to invariance violations is acceptable for their research question(s) and discuss the possible consequences when reporting their findings (Borsboom, 2006).

For further research, we have the following suggestions. First, we suggest to investigate whether the degree of LMI differs between specific subgroups, which may help explain the results. For example, Fokkema et al. (2013) found that LMI in the Beck Depression Inventory (Beck & Beamesderfer, 1974) was weaker for patients who received psychotherapy than for those who only received medication and additional clinical management. The authors suggested that less invariant measurement may be found in patients undergoing psychological treatments for depression due to a larger focus on the psychoeducation of patients. Thus, by studying specific subgroups, the authors found differences in the degree of LMI, and generated a hypothesis that may be studied further to possibly explain these differences. For more information on possible explanations for a lack of LMI, see Fried et al. (2016).

Second, it may be recommended to investigate whether modifications of the item banks will increase the degree of LMI. Specifically, it may be recommended to investigate the removal of items as rewriting or replacing them would be more complicated considering the comprehensive process of PROMIS to establish their item banks (Pilkonis et al., 2011). However, we should again stress that detecting individual noninvariant items is complex and many procedures are, to some extent, conceptually or statistically flawed. For example, Borsboom (2006) showed that using different methods for detecting noninvariant items can lead to different results. Also, researchers should realize that modifying an item bank, even when it concerns only one item, may lead to changes in the construct it measures. As a result, the set of items that shows invariance violations may change too (i.e., items that first showed sufficient invariance may found to be noninvariant for the modified item bank, and vice versa; Bechger & Maris, 2015). Furthermore, removing items could adversely affect content validity, and it can even result in more biased change scores because the equilibrium of biasing effects needed for cancellation to occur is disturbed (Borsboom, 2006). For these reasons, caution is warranted when item banks are modified. Alternatively, detecting individual noninvariant items may help to generate hypotheses about the origin of noninvariance. For example, it can be noted in our study that the individual items that showed the largest LMI violations assess anxiety very broadly (e.g., item EDANX05, *I felt anxious* or EDANX07 *I felt like I needed help for my anxiety*). This might imply that the anxiety construct as measured by the item bank actually consists of multiple constructs (e.g., generalized anxiety, social anxiety, and panic). In this case, bias may occur because patients think of different types of anxiety at separate test-occasions.

Third, we suggest studying LMI in patients with primary diagnoses other than anxiety or depression (e.g., attention deficit disorder, somatoform disorder or personality disorder), as the item banks also bear relevance for these patients. The reason for this is that depression and anxiety are often comorbid conditions (e.g., Löwe et al., 2008). Furthermore, anxiety and depression constitute a prime element of the distress that causes patients to seek help from mental health care professionals, also when their primary diagnosis is for instance a personality disorder (Leyro, Zvolensky, & Bernstein, 2010). In addition, we suggest that LMI is studied in populations without mental health problems, populations not in treatment, and general populations. Although changes in the observed item responses are expected to be low in these

populations, it is still fairly unclear what causes a lack of LMI (Fried et al., 2016). Therefore, the assumption of sufficient LMI in populations that do not show a substantial change in severity level over time should be studied.

Fourth, although the current study used a methodology that is the state of the art, additional new methods and software implementations would be welcome to study LMI in more detail. For example, LMI was evaluated in this study within the framework of factor analysis. In this framework, new methodology is available to investigate LMI for multiple group models that may also be extended to longitudinal models (Wu & Estabrook, 2016). Furthermore, although we investigated LMI with factor analysis because all new methodologies used in this study were primarily developed for this framework (Liu et al., 2017; Liu & West, 2018), PROMIS instruments are commonly calibrated using IRT, as it allows for the implementation of CAT. Studying equivalent longitudinal methods based on IRT (Meade & Lautenschlager, 2004; Wang, 2016) would allow for relating LMI violations to the metric used in clinical practice and the established properties of the item banks (Flens et al., 2017, 2019). A third example of new methodology concerns missing data. In the used version of the R package lavaan (i.e., 0.5-18), missing data handling is not available for CFA with ordered-categorical data (i.e., it uses listwise deletion). As missing data is common in longitudinal data, developing new methods that can handle missing data may result in improved parameter estimates.

In addition, the effect sizes used in this study were selected because, together, they provide highly practical information about the indicators of interest for test users (Liu & West, 2018). Specifically, they do not only provide information about the impact of invariance violations on change scores, but also on specific test-occasions, items, and response categories. However, the used rules of thumb for these effect sizes need to be verified in a (simulation) study to assess whether they correspond sufficiently to the proposed degree of bias. Furthermore, other effect sizes may provide additional useful information for test users (e.g., Choi, Gibbons, & Crane, 2011; Kim, Cohen, Alagoz, & Kim, 2007; Liu & West, 2018; Meade, 2010). A comparative (simulation) study on effect sizes and their rules of thumb used to quantify LMI with ordered-categorical indicators and for different applications of the item banks (e.g., full item bank administration, short-form administration, or CAT administration; Reeve et al., 2007) could provide new insights on the matter. In such a study, it could also be assessed whether the effect sizes could be further developed for evaluating LMI in individuals as compared with groups. Borsboom (2006) argued that when instruments are used for assessing individuals, LMI should conform to higher standards because of the increased danger of bias.

Fifth, we suggest to compare the degree of LMI between (a) the item banks and other instruments measuring Depression or Anxiety (e.g., the Center for Epidemiological Studies Depression scale or the Patient Health Questionnaire; Pilkonis et al., 2011) and (b) different languages (e.g., English and Dutch). By performing a comparative LMI study between instruments, test users have more available information to decide which instrument they want to use. Furthermore, it may provide new insights in the type of items that influence the degree of LMI. In addition, by performing a comparative LMI study between different languages, it could be assessed whether the lack of LMI may (also) be a translation problem.

In addition to the evaluation of LMI, the PROMIS item banks need to be studied on their responsiveness. According to the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) terminology (Mokkink et al., 2018), responsiveness (also known as sensitivity to change) refers to the ability to detect change in the measured construct over time (Mokkink et al., 2010), usually assessed by comparing changes in PROMIS scores to changes in one or several legacy instruments. Preferably, responsiveness should be studied for CAT administration rather than full item bank administration, as CAT will likely be the primary mode of administration in Dutch clinical practice. Moreover, we suggest to consider the results of the present study when comparing the responsiveness of the CAT administrations to that of other instruments. With CAT, the number of administered items is generally lower than with a full item bank administration. As a result, bias may be larger than in a full item bank administration as the items have a larger weight in the final test scores, and cancellation of biasing effects is less likely to occur (Borsboom, 2006).

In addition to responsiveness, we suggest to study whether multidimensional computerized adaptive testing (MCAT; Paap, Born, & Braeken, 2019) with the Depression and Anxiety item banks can be more efficient and precise than CAT based on separate unidimensional item banks. In the current study, the item banks were treated as measurements of separate unidimensional constructs because PROMIS deliberately chose to develop their instruments in this way (Cella et al., 2007). Numerous studies, however, show that the constructs depression and anxiety are highly correlated (e.g., de Beurs et al., 2007). Therefore, a logical next step with the PROMIS item banks could be to assess whether MCAT can be applied to the item banks. If this is the case, then LMI should once more be assessed for the multidimensional construct.

A strength of the current study is that the ordered-categorial data of the PROMIS item banks were explicitly treated as ordered-categorial instead of continuous, the latter being usually the case in LMI studies (Liu et al., 2017). Consequently, the item parameters may be more accurate (Rhemtulla et al., 2012). We also used two analyses to study the practical significance of the invariance violations, meaning that we gained information on (a) when (i.e., which test-occasion) and where (which item and response category) the problem occurred and (b) the magnitude of the problem for the parameter of interest in clinical practice (i.e., the mean latent change score; Liu & West, 2018). Finally, the patients' diagnoses were based on a standardized diagnostic interview (i.e., the MINI-plus; Sheehan et al., 1998), which will likely have increased the accuracy of the diagnoses compared with merely using the clinician's point of view (Aboraya, Rankin, France, El-Missiry, & John, 2006). However, although the MINI(-plus) has adequate diagnostic properties, studies did not show sufficient interrater agreement with other diagnostic instruments on detecting a generalized anxiety disorder and a simple phobia (Lecrubier et al., 1997; Sheehan et al., 1998). This may lead to underestimation or overestimation of these diagnoses. Overestimation may be unlikely, as the condition of each patient was deemed sufficiently severe to receive treatment. Underestimation may lead to these disorders being somewhat underrepresented in the present study sample.

In addition, there are several other reasons why the used samples of this study might lack representativeness for the Dutch clinical population. First, although we found that the

included patients did not differ substantially from the nonincluded patients in terms of gender, age distributions, and pretest score distributions, we could not evaluate the representativeness of the samples in terms of other variables that may affect LMI, such as type of treatment, comorbidity, or personality traits (e.g., agreeableness). We suggest to include these variables in future LMI studies. Moreover, the data should preferably be collected using stratified sampling (e.g., using stratification variables such as gender, age, education, ethnicity, and region; Flens et al., 2017). Second, we could not assess whether the change score distributions of the final samples were representative for the Dutch clinical population. It may be, for example, that patients with small change scores were more likely to refuse the posttest invitation because they did not respond to treatment. However, if such selection would be at play, it would hard if not impossible to overcome as participation in research is always voluntary. Third, the group who responded to the pretest, but were not invited for the posttest may appear large (Depression item bank, n = 1,470; Anxiety item bank, n = 951). According to Parnassia Psychiatric Institute (i.e., the mental health care provider that collected the data), reasons for this are diverse. For example, respondents could have dropped out of treatment (e.g., due to long waiting times or spontaneous remission), respondents' diagnoses could have changed during treatment, or treatment could have been terminated before the posttest was administered. As we did not know the specific reason for each individual that was not invited for the posttest, it is difficult to elaborate on how these reasons may have affected the representativeness of the samples for the Dutch clinical population. It may therefore be recommended that future studies administrate more specifically why respondents are not included in the study, but that may require a substantial investment.

The lack of uniformity in pretest to posttest interval could also have affected the results. To investigate the impact of this lack of uniformity on LMI to at least some extent, we repeated our analyses (not shown herein) on a more homogeneous subsample with additional inclusion criteria: (a) the pretest was administered before or on the day of the first treatment session and (b) the pretest and posttest were separated no longer than 12 months (Depression, n = 488; Anxiety, n = 414). We found that the results were highly similar, which can be seen as some evidence that the pretest to posttest interval is not a highly relevant factor in the degree of LMI. We should note, however, that these findings do not imply that the results would also have been highly similar when uniformity in the pretest to posttest interval was even larger (e.g., the pretest was administered at the first treatment session and the posttest exactly six months later). Unfortunately, we could not apply this larger extent of uniformity because the diminished sample size may result in data that is prone to nonconvergence, improper factor solutions, large standard errors, biased estimates of factor loadings and thresholds, and problematic goodness-of-fit tests (Liu et al., 2017). For future longitudinal studies, we suggest aiming for a higher degree of uniformity in the pretest to posttest interval to investigate more specific hypotheses about the length of the retest interval and LMI.

In addition to this, we suggest to evaluate LMI in more than two test-occasions. By investigating more test-occasions, the results may lead to a better understanding of the causes of invariance violations (e.g., by studying hypotheses concerning the impact of the degree of change on LMI). Extending the analyses of this study to more test-occasions is fairly straightforward. For an illustration of the analyses concerning the tenability of equality

constraints and the source of the invariance violations (i.e., which test-occasion, item, and response category), see Liu et al. (2017; 4 test-occasions). For an illustration of the analyses concerning the degree of impact on change scores, see Liu and West (2018; 4 test-occasions). Alternatively, the data sets used in this study could have been split into separate samples (e.g., a short-term and a long-term test-retest interval sample) to study LMI hypotheses (e.g., the effect of remembering items on LMI). However, we did not apply this approach because, again, the diminished sample size may result in data that is prone to nonconvergence, improper factor solutions, large standard errors, biased estimates of factor loadings and thresholds, and problematic goodness-of-fit tests (Liu et al., 2017).

Finally, the order of administered questionnaires at the pretest may have influenced the degree of LMI. This measurement was, in contrast to the posttest, preceded by several other questionnaires depending on a patient's disorder and age. Consequently, patients may have responded differently to items than they would have done when the PROMIS measurements were administered first (e.g., because of tiredness, or context effects; Windle, 1954).

In this study, we evaluated LMI in the Dutch-Flemish PROMIS item banks for Depression and Anxiety. Using tentatively determined rules of thumb, the results suggest that, even though some statistically significant violations of LMI were found, the item banks provide sufficiently invariant latent factor scores for use in clinical practice. This conclusion is often assumed for other (PROMIS) measurements. By assuming sufficient LMI, however, test users may have to deal with biased change scores without being aware of it. We therefore urge other researchers to study LMI in their own measurements.