CHAPTER 11
LONGITUDINAL MEASUREMENT INVARIANCE OF THE BECK SCALE FOR SUICIDE IDEATION

Published as:
Abstract

In mental health care, both clinical and scientific decisions are based on within-subject comparisons of test scores on the same self-report questionnaire at different points in time. To establish the validity of test score comparisons over time, longitudinal measurement invariance should be established. The current study tested whether the 19 item Beck Scale for Suicide Ideation is measurement invariant over time. As the first 5 items of the scale are often used to screen for the presence of suicidal thoughts, we also tested a model consisting of only the first five items. Psychiatric in- and out-patients (n = 475) completed the questionnaire upon admission and after three months. By means of confirmatory factor analysis we tested whether the parameters of a single factor model were equal over time. All fit indices indicated that both the 19-item questionnaire and the 5-item screener were measurement invariant over time. This means that changes in test-scores over time can be attributed to true changes in the construct of interest. These findings legitimate the use of the 19 item scale and the 5-item screener in longitudinal assessments.

Keywords

Suicide ideation, Measurement invariance, Response shift, Response bias, Screener

Introduction

After the start of the financial crisis in 2008, suicide rates increased substantially both in Europe and America 1, 2. A recent study showed 4884 additional suicides compared to the expected number based on the trend in 2000-2007 worldwide 1. Since 2007, in the Netherlands, the absolute number of suicides rose from 1353 to 1753, an increase of 30% 3. Governments are seeking new ways to improve care for suicidal patients 4, 5. Research on the effects of prevention strategies for suicide ideation may be used to improve health care 5. In both mental health care practice and research, self-reporting tools are used to assess and monitor patients’ health. Although a wide range of suicide scales are available, the Beck Scale for Suicide Ideation (BSS) is one of the most frequently used self-reporting scales for the assessment of suicidal thoughts 6, 7. It consists of 19 items and was developed to detect current intensity of a patient’s attitudes, plans and behaviors towards suicide. It also contains two additional items that assess the number of previous attempts and the intensity of the strength of the intent to die during the last attempt. The first five items serve as a screener for suicide ideation. More specifically, if a participant answers item 4 and 5 with the zero statements (indicating no active suicide intention and an avoidance of death if presented with a life-threatening situation), then he/she is allowed to skip all 14 additional items and asked to answer two extra items on previous suicide attempts. In common research and clinical practice, all first 5 items are used as a screener for suicide ideation 8, 9. Internal reliability, test-retest stability and concurrent validity for the BSS have been established in earlier studies 6, 10. The BSS is modeled on the basis of the interviewer-rated Scale for Suicide Ideation, one of the few suicide assessment tools with documented predictive validity for completed suicide 8.

In mental health care, a single administration can be used to provide information on the current status of the patient. But most of the time clinical and scientific decisions are based on the comparison of inter-subject scores on the same questionnaire at different points in time. The measurement over time is getting more important as policy makers and insurance companies ask for evidence of treatment effectiveness 11, 12. In Dutch mental health care, mental health institutions are required by insurance companies to assess patients’ mental health at regular intervals 13. Failure to do so leads to budget-cuts. Thus, clinical, scientific and management decisions are at least partly based on the change in scores on self-reporting measures over time. However, by their very nature, self-reporting measures are subjective 14. As a consequence, a patient’s understanding of the underlying construct measured by a questionnaire can change over time. For example, as psycho-education is recommended in most treatment guidelines 15, a patient’s
perception of the construct measured at different occasions, may change over time, due to treatment. Also, other changes in patient’s perceptions and internal standards may confound measurements over time.

**Measurement invariance and response shift**

To reach valid conclusions on the basis of repeated measurements, it has to be ensured that a patient’s score at the baseline measurement represents the same construct as the patients score at follow-up. In the literature on measurement of change, a distinction is made between measurements of real change and measurements confounded by changes in patients’ perception over time. A scale is thought to be measurement invariant (MI) if it measures the same construct across populations, groups or across measurement intervals. Most often, MI is studied across groups, to compare measurement models between for example different countries or ethnicities. However, MI can also be used to study measurement models within the same sample over time: longitudinal MI. This form of MI can be used to test for response shifts. The concept of response shift was introduced at the same time in education and management science. The following definition was provided.

*Response shift refers to a change in the meaning of one’s self-evaluation of a target construct as a result of a) a change in the respondent’s internal standards of measurement (scale recalibration); b) a change in the respondent’s values (reprioritization); or c) a redefinition of the target construct (reconceptualization) (p. 1532).*

To illustrate, imagine the same patient completing the BSS at two different times: once upon admission and once after three months of therapy. True change would mean an actual improvement or worsening of suicide ideation. Recalibration would mean that a patient has revised the response scale values between baseline and follow-up assessment. After therapy, a score of 1 on item 7 (I frequently think about killing myself) may reflect another level of rumination about suicide than before treatment. Reprioritization reflects the importance of an item in the context of the total scale. The writing of a suicide note (item 17) may have had more importance for a patient at baseline when compared to frequency of suicidal thoughts (item 7), but due to therapy, this prioritization may have changed. Finally, reconceptualization indicates a change in meaning of the content of the item. A patient’s understanding of his/her actual wish to die (item 2) may be different upon admission, than after three months of treatment, due to therapeutic intervention.

Following the operationalization by Oort (2005), the occurrence of response shift can be tested by comparing the factorial models that underlie consecutive assessments. More specifically, via confirmatory factor analysis (CFA), it can be tested whether the parameters of a factor model are different across consecutive measurements. Differences in item intercepts would indicate recalibration and differences in factor loadings would suggest reprioritization. Different salient factor loadings over time would indicate the occurrence of response shift due to reconceptualization.

To the authors’ knowledge, there has been no prior research into the (longitudinal) measurement invariance of the Beck Scale for Suicide Ideation (BSS). In the present study we have used the dataset from a multicenter controlled study to assess longitudinal MI of the 19-item BSS. As researchers and clinicians often use the first five BSS items to screen for suicide risk, we also tested longitudinal measurement invariance for the BSS using a model consisting only of the first five BSS items.

**Methods**

**Data set**

We used the data of the PITSTOP suicide study, a multicenter controlled study measuring the effect of guideline implementation on suicide ideation. Although the intervention was aimed at the training of professionals, the primary outcome was at patient level. Patients were assessed directly upon admission to the psychiatric department (T0) and after three months (T1). All patients were informed about the study and provided written informed consent before joining the study.

**Translation procedure**

The original version of the BSS was translated into Dutch for a clinical trial using the following process based on the WHO translation protocol: 1) Two PhD candidates, familiar with the terminology used in the scale, independently translated the original version of the BSS into Dutch. Both were fluent in English, but their primary language was Dutch. 2) Both translations were compared to create a consensus version 3) the consensus version was then translated back into English by an independent translator whose primary language is English and who had no knowledge of the questionnaire 4) the final version was reviewed by an expert in suicide prevention (Professor Kerkhof) and final adjustments were made. This final version was recently used in a trial.
Beck Scale for Suicide Ideation

The original BSS was developed in 1988, and was modeled after a successful interviewer-rated version, the Scale for Suicide Ideation. The BSS contains 19 items that measure the severity of actual suicidal wishes and plans. Scores range from 0-38, a higher score indicating a higher level of suicide ideation. Two studies indicated that the best cut-off to indicate high/low risk was BSS ≥ 2. Originally, if a patient scored 0 on item 4 and 5, the patient was directed to item 20. If the patient scored > 0 on item 4 and 5, all items of the BSS are completed. However, in most studies, the first 5 items are used as the screener. In the PITSTOP suicide study, participants were instructed to complete the full 19 items, even when they scored 0 on the first five items. In the current study, we included all patients that completed at least the first five items at both assessments and that completed a minimum of 2 items from items 6 through 19. The overall score is computed by totaling up the scores of the first 19 items.

Participants

For our analysis we used data from 872 patients who responded to baseline assessment, and 487 that completed 3 months follow-up. The preferred mode of data collection among patients was via the routine outcome monitoring system (ROM), an online system by which data on the effectiveness of treatment in everyday clinical practice are systematically collected. After the start of the study, it appeared impossible for most departments to collect our data via the ROM. In total, data of 287 (32%) patients was collected using the ROM. The rest of the data was collected via research assistants and clinicians that administered a paper and pencil version of the BSS to patients. Participants were instructed to complete the full BSS irrespective of scores on the first 5 items. However, to reduce patient burden, when a patient’s cumulative score was 0 on the first 5 items, and the patient did not want to continue answering questions, they were allowed to quit item administration after item 5. A total of 151 participants at baseline and 138 participants at T1 answered the first five items with 0 and then stopped. Following recommendations of Beck (1988), the remaining items were scored as 0. After scoring the remaining items with 0, missing items at both T0 and T1 of all other cases was 4%, which is negligible. As this is a pragmatic solution, we will repeat the analysis on the item score of the patients that completed all items at both occasions (n= 219), and on the dataset without the added zero scores on items 6-19. Of the 872 patients included at baseline, 457 (53%) were female. Average age was 43 (SD = 15), 567 (64%) scored BSS > 0 and 250 patients (30%) reported a history of suicide attempt. We were able to collect DSM-IV diagnoses for 549 (62%) of the 872 patients. Of these patients, 222 (40%) had a primary diagnosis of depression, 77 (14%) had a personality disorder, and 51 (9%) had a psychotic disorder.

Software

All analyses were performed in R (R development core team, 2009). All confirmatory factor analysis (CFA) models were estimated using the package lavaan, version 2.15.3 (Rosseel, 2012).

Estimation

Because of the ordinal nature of the data, mean- and variance-adjusted weighted least squares (WLSMV) estimation was used. In lavaan, this means that diagonally weighted least squares (DWLS) is used to estimate the model parameters, but the full weight matrix is used to compute robust standard errors, and a mean- and variance-adjusted test statistic. WLSMV estimation has been shown to result in unbiased parameter and standard error estimates, and acceptable type-I error rates for structural equation modeling with (skewed) ordinal variables.

Model identification

With WLSMV estimation of ordered categorical data, it is assumed that underlying every categorical observed response variable, there is a continuous latent response variable (i.e., observed item response variables are categorized continuous variables). These continuous latent response variables are regressed on a common (latent) factor. The scales of the latent variables are unknown, and have to be identified by fixing a number of parameters in the model. In the current study, the latent response variables were scaled to have unit variance. The variances of the common factors were identified by fixing the loading of the first item to one (at both time intervals). The mean structure of the model was identified by fixing all intercepts of the regressions of the latent response variables on the common factors, and the thresholds of the first response category of the first item to zero (at both time intervals).

Longitudinal measurement invariance

Most publications on MI use a multi-group CFA framework. Because the current study deals with longitudinal MI, a single group CFA framework was used, in which some variables in the model are allowed to correlate over time intervals, to take the longitudinal nature of the data into account.
The responses on the 19 BSS items within each measurement occasion were regressed on two common factors: suicidal ideation at T0 and suicidal ideation at T1. The common factors were allowed to correlate across time intervals. Similarly, residuals of the same continuous latent response variables were allowed to correlate between time intervals (Figure 1).

Figure 1: model to estimate measurement invariance

To assess MI of ordinal measures, three nested models are fit to the same dataset. First, a configural invariance model is estimated, in which loadings and thresholds are free parameters (with exception of parameters fixed for identification of the model). Second, a loading invariance model is estimated, in which the loadings are constrained to be equal across time intervals. Third, a threshold invariance model is estimated, in which the loadings, as well as the thresholds are constrained to be equal across time intervals. For every model, parameter estimates and model fit indices are inspected to evaluate model fit. Acceptable model fit of a less restricted model is a necessary condition for application of further model restrictions.

Model fit assessment

Overall model fit can be assessed by means of model fit indices. The use of several fit indices for evaluating model fit is recommended, to minimize type I and type II error. In the current study, root mean square error of approximation (RMSEA), comparative fit index (CFI), and the minimum function test statistic are used. For CFI, models with values ≥ 0.95 have acceptable fit. For RMSEA, models with values ≤ 0.06 have acceptable fit. Finally, the minimum function test statistic can be evaluated by comparing its value to a chi-square distribution with degrees of freedom (df) equal to the df of the model. However, because of their dependency on sample size (i.e., they are almost always significant with N ≥ 400), we have focused on assessing model fit with CFI and RMSEA.

In addition, the fit of two nested models can be compared by taking the difference of the fit indices. For WLSMV estimation, a scaled chi-square difference test for nested models can be computed in lavaan. However, the scaled chi-square difference suffers from the same dependency on sample size as the minimum fit function statistic. Therefore, we focused on changes in model fit according to CFI and RMSEA. We used the criteria suggested by Chen (2008): a decrease in CFI of ≥ 0.01, and an increase in RMSEA of ≥ 0.015 was taken as an unacceptable decrease of model fit. Chen (2008) suggested a cutoff of 0.005 for sample sizes < 300, uniform patterns of non-invariance and unequal sample sizes. Meade, Johnson and Braddy (2008) suggested stricter criteria: a decrease in CFI of ≥ 0.002 should be taken as an unacceptable decrease of model fit.

5-item screener

To test longitudinal measurement invariance for the 5-item screener, we used a model consisting only of the first five BSS items. The methods to analyze MI were the same as described above for the 19-item scale. Results of the 5-item screener were presented separately.
Results

Descriptives

Table 1 presents the descriptive statistics of the items scores at baseline and follow up. The mean of all individual items show a decrease from T0 to T1, indicating that at T1, patients show lower scores on the suicide ideation scale than at baseline. Cronbach’s alphas were similar at T0 (0.93) and at T1 (0.93).

Table 1: descriptives of BSS-NL item scores at T0 and T1

<table>
<thead>
<tr>
<th>#</th>
<th>content</th>
<th>T0</th>
<th></th>
<th>T1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>wish to live</td>
<td>868</td>
<td>0.50</td>
<td>487</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>wish to die</td>
<td>866</td>
<td>0.65</td>
<td>486</td>
<td>0.40</td>
</tr>
<tr>
<td>3</td>
<td>reasons living/dying</td>
<td>870</td>
<td>0.46</td>
<td>487</td>
<td>0.26</td>
</tr>
<tr>
<td>4</td>
<td>desire to kill oneself</td>
<td>872</td>
<td>0.51</td>
<td>487</td>
<td>0.31</td>
</tr>
<tr>
<td>5</td>
<td>save my life</td>
<td>851</td>
<td>0.53</td>
<td>487</td>
<td>0.31</td>
</tr>
<tr>
<td>6</td>
<td>periods thinking about suicide</td>
<td>797</td>
<td>0.41</td>
<td>453</td>
<td>0.23</td>
</tr>
<tr>
<td>7</td>
<td>frequency of thinking about suicide</td>
<td>800</td>
<td>0.37</td>
<td>451</td>
<td>0.22</td>
</tr>
<tr>
<td>8</td>
<td>acceptance of idea of suicide</td>
<td>816</td>
<td>0.46</td>
<td>461</td>
<td>0.32</td>
</tr>
<tr>
<td>9</td>
<td>ability to not commit suicide</td>
<td>819</td>
<td>0.29</td>
<td>461</td>
<td>0.16</td>
</tr>
<tr>
<td>10</td>
<td>reasons for not committing suicide</td>
<td>807</td>
<td>0.36</td>
<td>465</td>
<td>0.25</td>
</tr>
<tr>
<td>11</td>
<td>reasons for wanting to commit suicide</td>
<td>751</td>
<td>0.89</td>
<td>424</td>
<td>0.63</td>
</tr>
<tr>
<td>12</td>
<td>specific plan to commit suicide</td>
<td>795</td>
<td>0.41</td>
<td>448</td>
<td>0.29</td>
</tr>
<tr>
<td>13</td>
<td>access to suicide method</td>
<td>742</td>
<td>0.49</td>
<td>441</td>
<td>0.40</td>
</tr>
<tr>
<td>14</td>
<td>courage/ability to commit suicide</td>
<td>800</td>
<td>0.48</td>
<td>451</td>
<td>0.32</td>
</tr>
<tr>
<td>15</td>
<td>expectance to commit suicide</td>
<td>809</td>
<td>0.35</td>
<td>460</td>
<td>0.20</td>
</tr>
<tr>
<td>16</td>
<td>preparations for suicide</td>
<td>791</td>
<td>0.20</td>
<td>455</td>
<td>0.12</td>
</tr>
<tr>
<td>17</td>
<td>writing of suicide note</td>
<td>776</td>
<td>0.24</td>
<td>448</td>
<td>0.15</td>
</tr>
<tr>
<td>18</td>
<td>arrangements for after suicide</td>
<td>789</td>
<td>0.3</td>
<td>452</td>
<td>0.28</td>
</tr>
<tr>
<td>19</td>
<td>conceal ideation</td>
<td>771</td>
<td>0.6</td>
<td>457</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Note. n = number of participants, M = mean, VAR = variance

Invariance models for the 19-item scale

Table 2 presents the fit indices for each of the invariance models. The configural invariance model proved to have good model fit, judging by its CFI value of 0.981 and RMSEA value of 0.041. In addition, the 90% confidence interval of the RMSEA did not include 0.06. However, a negative error variance for item 13 was observed at T1, although the model estimation converged normally, and all other parameter estimates had plausible values. Following recommendations (Chen et al., 2001), we constrained the error covariance across time intervals for item 13 to a theoretically plausible value. In the unconstrained model, the error variance of item 13 was 0.390 at T0, and 0.158 at T1; so theoretically, the error covariance cannot exceed √(0.390·0.158) = 0.248. In the unconstrained model, the error covariance across time intervals for item 13 was 0.260 (SE = 0.044), resulting in a negative error variance for item 13 at T1. Therefore, we fixed the error covariance to 0.245 in the configural invariant model. This did not result in a change in model fit indices, with exception of a slight improvement of the minimum fit function value.

Restricting all loadings to be equal across time intervals resulted in a slight improvement in model fit, according to the increase in CFI of 0.001 and decrease in RMSEA of 0.001. No negative variances were observed in this model, so the error covariance across time intervals for item 13 was freely estimated in the models with invariant loadings.

Next, restricting all item thresholds to be equal across time intervals resulted in a slight improvement of model fit, as well. CFI increased by 0.001, RMSEA decreased by 0.001, and the entire 90% confidence interval of the RMSEA remained < 0.06.

Performing the analysis on the dataset of completers only, and the dataset without added zeros for items 6-19, resulted in very similar findings: RMSEA and CFI for the configural invariance model showed good model fit, and restricting loadings and thresholds to equality across time intervals resulted in slight improvements of model fit. In addition, parameter estimates for the configural invariance model were very similar for all three datasets.
Table 2. Model fit indices for the MI models

<table>
<thead>
<tr>
<th>Model</th>
<th>minimum fit function</th>
<th>df</th>
<th>p-value</th>
<th>scaled chi-square difference test statistic</th>
<th>df</th>
<th>p-value</th>
<th>CFI</th>
<th>RMSEA</th>
<th>RMSEA 90%-CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural invariance</td>
<td>1391.887</td>
<td>646</td>
<td>&lt;.0001</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>0.981</td>
<td>0.041</td>
<td>0.038-0.0443</td>
</tr>
<tr>
<td>Loading invariance</td>
<td>1368.792</td>
<td>663</td>
<td>&lt;.0001</td>
<td>50.275</td>
<td>17</td>
<td>&lt;.001</td>
<td>0.982</td>
<td>0.040</td>
<td>0.037-0.0433</td>
</tr>
<tr>
<td>Threshold invariance</td>
<td>1424.050</td>
<td>700</td>
<td>&lt;.0001</td>
<td>96.652</td>
<td>37</td>
<td>&lt;.001</td>
<td>0.982</td>
<td>0.039</td>
<td>0.036-0.0423</td>
</tr>
</tbody>
</table>

Note. All fit indices based on robust standard errors and WLSMV estimator. df = degrees of freedom, CFI = comparative fit index, RMSEA = root mean square error of approximation, CI = confidence interval

Invariance models for the 5-item screener

Table 3 presents the fit indices for each of the invariance models consisting only of the first 5 BSS items. It indicates a good fit for the unidimensional models on both measurement occasions: CFI was 0.997 and the 90% confidence interval of the RMSEA was <.06. Restricting factor loadings and thresholds to equality on both measurement occasions resulted in an improvement of fit, judging by the decrease in RMSEA, and no deterioration according to the CFI, which remained the same.

Discussion

The present study is the first to examine longitudinal measurement invariance of the Beck Scale for Suicide Ideation. Longitudinal measurement invariance is a prerequisite for the comparison of repeated measurements with a (mental health) test, such as when a test is used for routine outcome monitoring \(^8,9\). A lack of measurement invariance can lead to confounded interpretation of change scores. For example, a recent study found the widely used Beck Depression Inventory was not measurement invariant over the course of depression treatment \(^5\). This resulted in an underestimation of depressive symptoms at baseline compared to follow-up measurement. In addition, measurement errors were smaller, and correlations between different constructs of the BDI were stronger at follow-up. The consequence is that comparison of the observed total scores of the BDI may underestimate treatment efficacy and result in biased conclusions. Via a longitudinal CFA model we examined the occurrence of response shift when using the Beck Scale for Suicide Ideation in a multicenter RCT. Our results show the BSS to be measurement invariant over time. This means no relevant response shift occurred, and any change found over time on the scale can be interpreted as actual change. These findings legitimize the use of the scale in longitudinal assessments. One of the reasons that the BSS is invariant could be that the BSS is a unidimensional scale with very specific and clearly worded items such as: “I have the courage to commit suicide”. When compared to the three dimensional BDI, with more broad items like: “I am disappointed in myself”, the items of the BSS are more straightforward, and therefore less vulnerable to differing interpretations over time. Importantly, the first five items were also found to be measurement invariant. As asking about suicidality can be difficult for both patients \(^39,40\) and professionals \(^4\), a short screener that is measurement invariant over time is much desired. Our findings indicate that the 5-item screener of the BSS can be used over time, and that change on the 5-item screener represents true change in suicide ideation.

This study has several limitations. First, although patients and research assistants were encouraged to complete all 19 items even when a participant scored 0 on the first 5 items \(^7\), 22% (n = 188) of the patients at T0 and 33% (n = 180) at T1 stopped after answering the first 5 items with 0. Also, considering the participants that continued responding after item 5, 4% of the scores were missing. As is common when using the BSS \(^4\), we estimated the missing items to reflect score 0, but this may be an underestimation of actual suicide ideation. As the procedure for dealing with missing values was the same at baseline and at T1, and an additional completers-only analysis showed no differences in the values of fit indices, this should not have led to different conclusions concerning MI of the BSS.

Strength of the current study is the large number of psychiatric patients with two assessments of the BSS. Normally, the number of respondents to suicide questionnaires is too small for assessment of longitudinal MI \(^7\). Also, for reasons discussed elsewhere \(^4\), MI is seldom investigated in longitudinal mental health trials. By investigating MI in a large study population, we further validated the BSS and justified its longitudinal use. It is necessary to compare the CFA parameters of our study with, for example, data collected with the original English BSS. This study hopes to encourage other researchers in mental health care to test MI of frequently used mental health questionnaires, and thereby identify the measurement of true and potentially confounded changes before drawing clinical, managerial or scientific conclusions.
Abbreviations

MI = measurement invariance, CFA = confirmatory factor analysis, BSS = Beck Scale for Suicide ideation, CFI = comparative fit index, RMSEA = root mean square error of approximation, CI = confidence interval, ROM = routine outcome monitoring, PITSTOP suicide = professionals in training to STOP suicide, DF = degrees of freedom

Competing interests

All authors declare that they have no competing interests

Authors’ contributions

AK, MdG, and JdK obtained funding for this study. DdB carried out the study. DdB and MF drafted the manuscript. AK, MdG, JdK contributed to the execution of the study, and to the manuscript writing.

Acknowledgements

This study is funded by The Dutch organization for health research and development (ZonMW)

References


24. de Beurs DP, de Groot MH, de Keijser J, Twisk JW, van Duijn E, van Hemert AM, Verlinde L, Spijker10 J: Improving the application of a practice guideline for the assessment and treatment of suicidal behavior by training the full staff of psychiatric departments via an e-learning supported Train-the-Trainer program: study protocol for a randomized controlled BMC trials 2013, 14(9).


