Psychiatry Research 225 (2015) 368–373

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Longitudinal measurement invariance of the Beck Scale for Suicide Ideation

Derek P. de Beurs [a,b,*], Marjolein Fokkema [a,b], Marieke H. de Groot [a,b], Jos de Keijser [c], Ad J.F.M. Kerkhof [a,b]

[a] Department of Clinical Psychology, VU University, Amsterdam, The Netherlands
[b] EMGO Institute for Health and Care Research, Amsterdam, The Netherlands
[c] GGZ Foundation for Mental Health Care Friesland and Groningen University, Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 12 March 2014
Received in revised form
18 September 2014
Accepted 24 November 2014
Available online 23 December 2014

Keywords:
Suicide ideation
Measurement invariance
Response shift
Response bias
Screener

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 (BSS) is measurement invariant (MI) over time. As the first five 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 3 months. By means of confirmatory factor analysis (CFA) 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 five-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 five-item screener in longitudinal assessments.

1. Introduction

After the start of the financial crisis in 2008, suicide rates increased substantially both in Europe and America (Gunnell et al., 2009; Stuckler and Basu, 2013). A recent study showed 4884 additional suicides compared to the expected number based on the trend in 2000–2007 worldwide (Gunnell et al., 2009). Since 2007, in the Netherlands, the absolute number of suicides rose from 1353 to 1753, an increase of 30% (CBS, 2013). Governments are seeking new ways to improve care for suicidal patients (de Beurs et al., 2013a; Hegerl et al., 2013). Research on the effects of prevention strategies for suicide ideation may be used to improve health care (de Beurs et al., 2013a). 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 (Beck et al., 1988; Brown, 2001). 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 items 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 five items are used as a screener for suicide ideation (Brown et al., 2000; van Spijker et al., 2010). Internal reliability, test–retest stability and concurrent validity for the BSS have been established in earlier studies (Brown, 2001; De Beurs et al., 2014). 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 (Brown et al., 2000).

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; d.f., degrees of freedom

* Corresponding author at: Department of Clinical Psychology, VU University, Amsterdam, The Netherlands. Tel.: +31 6 24504141; fax: +31 20 3459201.
E-mail addresses: dp.de.beurs@vu.nl, derekdebeurs@gmail.com (D.P. de Beurs), m.fokkema@vu.nl (M. Fokkema), mariekedegroot@ziggo.nl (M.H. de Groot), jos.de.keijser@ggzfriesland.nl (J. de Keijser), ajfm.kerkhof@vu.nl (A.J.F.M. Kerkhof).

http://dx.doi.org/10.1016/j.psychres.2014.11.075
0165-1781/

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 (Van Os et al., 2012; Young et al., 2000). In Dutch mental health care, mental health institutions are required by insurance companies to assess patients' mental health at regular intervals (De Beurs et al., 2011). 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 (Borsboom, 2005). 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 (Lukens and McFarlane, 2004), a patient's perception of the construct measured at different occasions may change over time, due to treatment (Fokkema et al., 2013). Also, other changes in patient's perceptions and internal standards may confound measurements over time (Sprangers and Schwartz, 1999).

1.1. 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 (Vandenberg and Lance, 2000). A scale is thought to be measurement invariant (MI) if it measures the same construct across populations, groups or across measurement intervals (Millsap, 2011). Most often, MI is studied across groups, to compare measurement models between for example different countries or ethnicities (Vandenberg and Lance, 2000). 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 (Howard et al., 1979) and management science (Golembiewski et al., 1976). The following definition was provided (Schwartz and Sprangers, 1999):

> *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 3 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 3 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 (Oort, 2005).

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 (de Beurs et al., 2013a) 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.

2. Methods

2.1. Data set

We used the data of the PITSTOP suicide study (de Beurs et al., 2013a, 2013b), a multicenter controlled study measuring the effect of guideline implementation on suicide ideation. Although the intervention was aimed at the training of professionals (de Beurs et al., 2013b), the primary outcome was at patient level. Patients were assessed directly upon admission to the psychiatric department (T0) and after 3 months (T1). All patients were informed about the study and provided written informed consent before joining the study.

2.2. Translation procedure

The original version of the BSS (Beck et al., 1988) was translated into Dutch for a clinical trial (van Spijker et al., 2010) using the following process based on the WHO translation protocol (Demyttenaere et al., 2013): 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 (van Spijker et al., 2012, 2014).

2.3. 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 (Beck et al., 1979). The BSS contains 19 items that measure the severity of actual suicidal wishes and plans. Scores range from 0 to 38, a higher score indicating a higher level of suicide ideation. Two studies (Beck et al., 1999; Brown et al., 2000) indicated that the best cut-off to indicate high/low risk was BSS >2. Originally, if a patient scored 0 on items 4 and 5, the patient was directed to item 20. If the patient scored >0 on items 4 and 5, all items of the BSS are completed. However, in most studies, the first five items are used as the screener (Brown et al., 2000; van Spijker et al., 2010). In the PITSTOP suicide study, participants were instructed to complete the full 19 items, even when they scored 0 on the first five items (de Beurs et al., 2013a; De Beurs et al., 2014). 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 two items from items 6 through 19. The overall score is computed by totaling up the scores of the first 19 items.

2.4. 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 (De Beurs et al., 2011). 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. Patients were instructed to complete the full BSS irrespective of scores on the first five items. However, to reduce patient burden, when a patient's cumulative score was 0 on the first five 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 et al. (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 were 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 (S.D.=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.

2.5. 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).

2.6. Estimation

Because of the ordinal nature of the data, mean- and variance-adjusted weighted least squares (WLSMV) estimation was used (Reeve et al., 2007). In lavaan, this means that diagonally weighted least squares (DWLS) are 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 (Flora and Curran, 2004; Lei, 2009).

2.7. 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).

2.8. Longitudinal measurement invariance

Most publications on MI use a multi-group CFA framework (Vandenberg and Lance, 2000). 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 (Fokkema et al., 2013). 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 (Fig. 1).

To assess MI of ordinal measures, three nested models are fit to the same dataset (Millsap and Yun-Tein, 2004). 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.

2.9. 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 (Bentler, 1990). 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 (Hu and Bentler, 1999). For RMSEA, models with values ≤ 0.06 have acceptable fit (Hu and Bentler, 1999). Finally, the minimum function test statistic can be evaluated by comparing its value to a chi-square distribution with degrees of freedom (d.f.) equal to the d.f. of the model. However, because of their dependency on sample size (i.e., they are almost always significant with $N \geq 400$), we have focused on assessing model fit with CFI and RMSEA (Cheung and Rensvold, 2002).

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 (Satorra, 2000) 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 et al. (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 et al. (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.

2.10. Five-item screener

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

3. Results

3.1. 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).

3.2. 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 $\sqrt{(0.390 \bullet 0.158)} = 0.248$. 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.

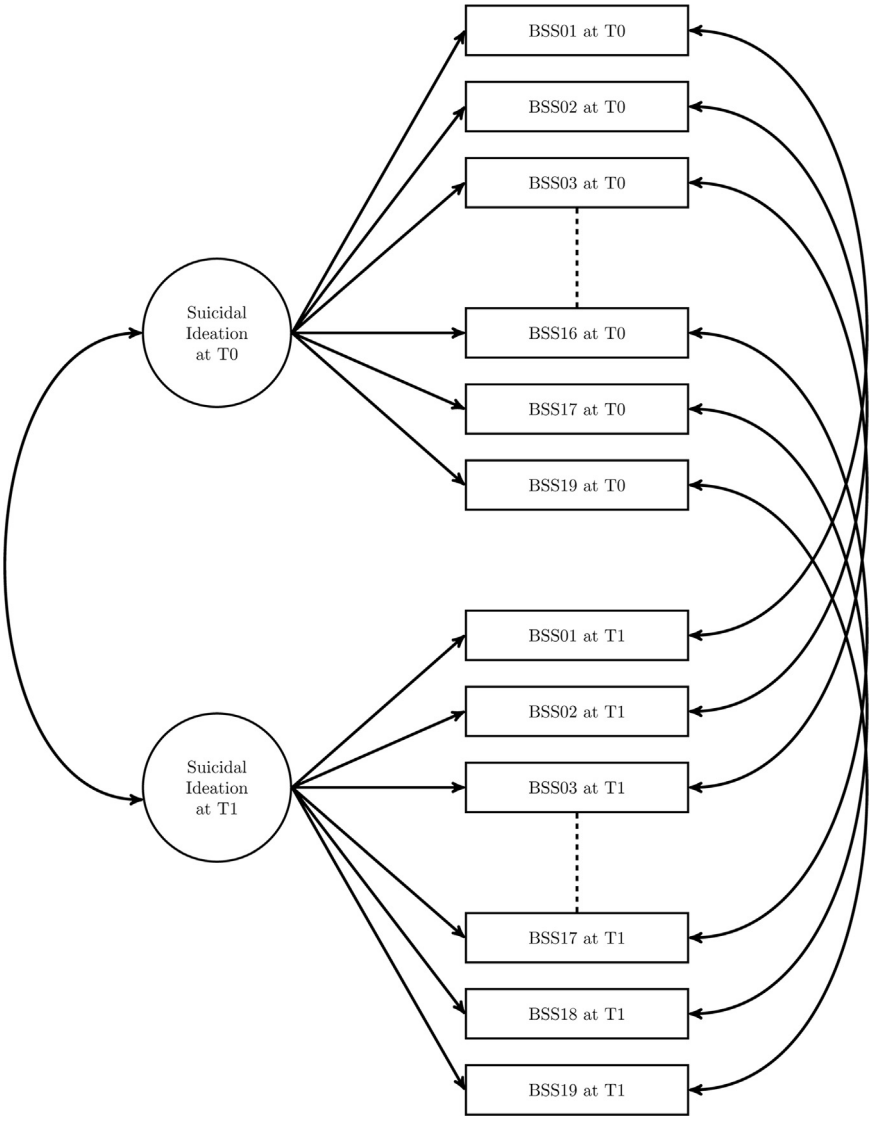

Fig. 1. Model to estimate measurement invariance.

Table 1
Descriptives of BSS-NL item scores at T0 and T1.

#	Content	T0			T1		
		n	*M*	VAR	*n*	*M*	VAR
1	Wish to live	868	0.50	0.84	487	0.28	1.41
2	Wish to die	866	0.65	0.63	486	0.40	1.06
3	Reasons living/dying	870	0.46	0.92	487	0.26	1.46
4	Desire to kill oneself	872	0.51	0.81	487	0.31	1.31
5	Save my life	851	0.53	0.79	487	0.31	1.29
6	Periods thinking about suicide	797	0.41	1.02	453	0.23	1.58
7	Frequency of thinking about suicide	800	0.37	1.14	451	0.22	1.63
8	Acceptance of idea of suicide	816	0.46	0.93	461	0.32	1.26
9	Ability to not commit suicide	819	0.29	1.36	461	0.16	1.87
10	Reasons for not committing suicide	807	0.36	1.15	465	0.25	1.51
11	Reasons for wanting to commit suicide	751	0.89	0.46	424	0.63	0.64
12	Specific plan to commit suicide	795	0.41	1.03	448	0.29	1.38
13	Access to suicide method	742	0.49	0.86	441	0.40	1.06
14	Courage/ability to commit suicide	800	0.48	0.88	451	0.32	1.28
15	Expectance to commit suicide	809	0.35	1.19	460	0.20	1.69
16	Preparations for suicide	791	0.20	1.69	455	0.12	2.06
17	Writing of suicide note	776	0.24	1.54	448	0.15	1.93
18	Arrangements for after suicide	789	0.3	1.26	452	0.28	1.42
19	Conceal ideation	771	0.6	0.93	457	0.32	1.28

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

Table 2
Model fit indices for the MI models.

Model	Minimum fit function	d.f.	*p*-Value	Scaled chi-square difference test statistic			CFI	RMSEA	RMSEA 90%-CI
				Test statistic	d.f.	*p*-Value			
Configural invariance	1391.887	646	< 0.001	N/A	N/A	N/A	0.981	0.041	0.038–0.044
Loading invariance	1368.792	663	< 0.001	50.275	17	< 0.001	0.982	0.040	0.037–0.043
Threshold invariance	1424.050	700	< 0.001	96.652	37	< 0.001	0.982	0.039	0.036–0.042

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

Table 3
Model fit indices for the five-item MI models.

Model	Minimum fit function	d.f.	*p*-Value	Scaled chi-square difference test statistic			CFI	RMSEA	RMSEA 90%-CI
				Test statistic	d.f.	*p*-Value			
Configural invariance	84.989	29	< 0.001	N/A	N/A	N/A	0.997	0.047	0.035–0.058
Loading invariance	89.403	33	< 0.001	10.806	4	0.029	0.997	0.044	0.033–0.055
Threshold invariance	95.638	42	< 0.001	2.9681	9	< 0.001	0.997	0.038	0.028–0.048

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

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.

3.3. *Invariance models for the five-item screener*

Table 3 presents the fit indices for each of the invariance models consisting only of the first five 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 < 0.06. 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.

4. 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 (De Beurs et al., 2011; Meredith, 1993). A lack of measurement invariance can lead to confounded interpretation of change scores. For example, a recent study found that the widely used Beck Depression Inventory was not measurement invariant over the course of depression treatment (Fokkema et al., 2013). 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 (Jorm et al., 2007; Younes et al., 2013) and professionals (Sher, 2011), a short screener that is measurement invariant over time is much desired. Our findings indicate that the five-item screener of the BSS can be used over time, and that change on the five-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 five items (de Beurs et al., 2013a), 22% ($n=188$) of the patients at T0 and 33% ($n=180$) at T1 stopped after answering the first five 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 (Brown et al., 2000; van Spijker et al., 2010), 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 (van Spijker et al., 2010). Also, for reasons discussed elsewhere (Borsboom, 2006), 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.

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.

Acknowledgments

This study is funded by The Dutch Organization for Health Research and Development (ZonMw grant: 171103006).

References

Beck, A.T., Brown, G.K., Steer, R.A., Dahlsgaard, K.K., Grisham, J.R., 1999. Suicide ideation at its worst point: a predictor of eventual suicide in psychiatric outpatients. Suicide and Life-Threatening Behavior 29, 1–9.

Beck, A.T., Kovacs, M., Weissman, A., 1979. Assessment of suicidal intention: the scale for suicide ideation. Journal of Consulting and Clinical Psychology 47, 343.

Beck, A.T., Steer, R.A., Ranieri, W.F., 1988. Scale for suicide ideation: psychometric properties of a self-report version. Journal of Clinical Psychology 44, 499–505.

Bentler, P.M., 1990. Comparative fit indexes in structural models. Psychological Bulletin 107, 238.

Borsboom, D., 2005. Measuring the Mind: Conceptual Issues in Contemporary Psychometrics. Cambridge University Press, New York.

Borsboom, D., 2006. The attack of the psychometricians. Psychometrika 71, 425–440.

Brown, G.K., 2001. A review of suicide assessment measures for intervention research with adults and older adults. Available at ⟨http://ruralccp.org/lyra-data/storage/asset/brown-nd-27cb.pdf⟩. Retrieved on 18 september 2014.

Brown, G.K., Beck, A.T., Steer, R.A., Grisham, J.R., 2000. Risk factors for suicide in psychiatric outpatients: a 20-year prospective study. Journal of Consulting and Clinical Psychology 68, 371.

CBS, 2013. Available at ⟨http://statline.cbs.nl/statweb⟩. Retrieved on 18 september 2014.

Chen, F., Bollen, K.A., Paxton, P., Curran, P.J., Kirby, J.B., 2001. Improper solutions in structural equation models. Causes, consequences, and strategies. Sociological Methods & Research 29, 468–508.

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

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

de Beurs, D.P., de Groot, M.H., Bosmans, J.E., de Keijser, J., Mokkenstorm, J., Verwey, B., van Duijn, E., de Winter, R.F., Kerkhof, A.J., 2013a. Reducing patients' suicide ideation through training mental health teams in the application of the Dutch multidisciplinary practice guideline on assessment and treatment of suicidal behavior: study protocol of a randomized controlled trial. Trials 14, 372.

de Beurs, D.P., de Groot, M.H., de Keijser, J., Verwey, B., Mokkenstorm, J., Twisk, J.W., van Duijn, E., van Hemert, A.M., Verlinde, L., Spijker, J., 2013b. 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 trial. Trials 14, 9.

De Beurs, D.P., de Vries, A.L., de Groot, M.H., de Keijser, J., Kerkhof, A.J., 2014. Applying computer adaptive testing to optimize online assessment of suicidal behavior: a simulation study. Journal of Medical Internet Research 16, e207.

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

Demyttenaere, K., Bruffaerts, R., Posada-Villa, J., Gasquet, I., Kovess, V., Lepine, J., Angermeyer, M.C., Bernert, S., de Girolamo, G., Morosini, P., 2013. Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization. Journal of the American Medical Association 291, 2581–2590.

Flora, D.B., Curran, P.J., 2004. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychological Methods 9, 466.

Fokkema, M., Smits, N., Kelderman, H., Cuijpers, P., 2013. Response shifts in mental health interventions: an illustration of longitudinal measurement invariance. Psychological Assessment 25, 520.

Golembiewski, R.T., Billingsley, K., Yeager, S., 1976. Measuring change and persistence in human affairs: types of change generated by OD designs. The Journal of Applied Behavioral Science 12, 133–157.

Gunnell, D., Platt, S., Hawton, K., 2009. The economic crisis and suicide. British Medical Journal 338, b1891.

Hegerl, U., Rummel-Kluge, C., Värnik, A., Arensman, E., Koburger, N., 2013. Alliances against depression—a community based approach to target depression and to prevent suicidal behaviour. Neuroscience & Biobehavioral Reviews 37, 2404–2409.

Howard, G.S., Ralph, K.M., Gulanick, N.A., Maxwell, S.E., Nance, D.W., Gerber, S.K., 1979. Internal invalidity in pretest–posttest self-report evaluations and a re-evaluation of retrospective pretests. Applied Psychological Measurement 3, 1–23.

Hu, L.t., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling 6, 1–55.

Jorm, A.F., Kelly, C.M., Morgan, A.J., 2007. Participant distress in psychiatric research: a systematic review. Psychological Medicine 37, 917–926.

Lei, P.W., 2009. Evaluating estimation methods for ordinal data in structural equation modeling. Quality and Quantity 43, 495–507.

Lukens, E.P., McFarlane, W.R., 2004. Psychoeducation as evidence-based practice: considerations for practice, research, and policy. Brief Treatment and Crisis Intervention 4, 205.

Meade, A.W., Johnson, E.C., Braddy, P.W., 2008. Power and sensitivity of alternative fit indices in tests of measurement invariance. Journal of Applied Psychology 93, 568.

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

Millsap, R.E., 2011. Statistical Approaches to Measurement Invariance. Routledge, New York.

Millsap, R.E., Yun-Tein, J., 2004. Assessing factorial invariance in ordered-categorical measures. Multivariate Behavioral Research 39, 479–515.

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

R development Core Team, 2009. R Project for Statistical Computing. Vienna, Austria.

Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., Thissen, D., Revicki, D.A., Weiss, D.J., Hambleton, R.K., 2007. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the patient-reported outcomes measurement information system (PROMIS). Medical Care 45, S22–S31.

Rosseel, Y., 2012. lavaan: an R package for structural equation modeling. Journal of Statistical Software 48, 1–36.

Satorra, A., 2000. Scaled and adjusted restricted tests in multi-sample analysis of moment structures. Advanced Studies in Theoretical and Applied Econometrics 36, 233–247.

Schwartz, C.E., Sprangers, M.A., 1999. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Social Science & Medicine 48, 1531–1548.

Sher, L., 2011. Teaching medical professionals about suicide prevention: what's missing? QJM 104, 1005–1008.

Sprangers, M.A., Schwartz, C.E., 1999. Integrating response shift into health-related quality of life research: a theoretical model. Social Science & Medicine 48, 1507–1515.

Stuckler, D., Basu, S., 2013. The Body Economic: Why Austerity Kills. Basic Books, New York.

Van Os, J., Kahn, R., Denys, D., Schoevers, R., Beekman, A., Hoogendijk, W., van Hemert, A., Hodiamont, P., Scheepers, F., Delespaul, P.A., 2012. ROM: gedragsnorm of dwangmaatregel? Overwegingen bij het themanummer over routine outcome monitoring. Tijdschrift voor Psychiatrie 54, 245–253.

van Spijker, B.A., Majo, M.C., Smit, F., van Straten, A., Kerkhof, A.J., 2012. Reducing suicidal ideation: cost-effectiveness analysis of a randomized controlled trial of unguided web-based self-help. Journal of Medical Internet Research 14 (5), e141.

van Spijker, B.A., van Straten, A., Kerkhof, A., 2010. The effectiveness of a web-based self-help intervention to reduce suicidal thoughts: a randomized controlled trial. Trials 11, 25.

van Spijker, B.A., van Straten, A., Kerkhof, A.J., 2014. Effectiveness of online self-help for suicidal thoughts: results of a randomised controlled trial. PLoS One 9, e90118.

Vandenberg, R.J., Lance, C.E., 2000. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organizational Research Methods 3, 4–70.

Younes, N., Chee, C.C., Turbelin, C., Hanslik, T., Passerieux, C., Melchior, M., 2013. Particular difficulties faced by GPs with young adults who will attempt suicide: a cross-sectional study. BMC Family Practice 14, 68.

Young, A.S., Grusky, O., Jordan, D., Belin, T.R., 2000. Routine outcome monitoring in a public mental health system: the impact of patients who leave care. Psychiatric Services 51, 85–91.