Martin, C.R., Vardavaki, Z. and Hollins Martin, C.J. (in press). Measurement equivalence of the Birth Satisfaction Scale-Revised (BSS-R): Further evidence of construct validity. *Journal of Reproductive and Infant Psychology.*

Measurement equivalence of the Birth Satisfaction Scale-Revised (BSS-R): further evidence of construct validity

Colin R.Martin[1]

Zoi Vardavaki[2]

&

Caroline J. Hollins Martin[3]

[1]Professor of Mental Health, Faculty of Society and Health, Buckinghamshire New University, Uxbridge, UK, UB8 1NA. E-mail: colin.martin@bucks.ac.uk

[2]Midwife in University College London Hospitals NHS Foundation Trust, London, NW12PG, UK;, UK. E-mail: zoevardavaki@gmail.com

[3]Professor in Maternal Health, School of Nursing, Midwifery and Social Care Edinburgh Napier University, Scotland, UK. EH11 4BN.
E-mail: C.HollinsMartin@napier.ac.uk

Address for correspondence

Address for correspondence: Professor Colin R Martin, Room 2.11, Faculty of Health and Society, Buckinghamshire New University, Uxbridge Campus
106 Oxford Road, Uxbridge, Middlesex, UB8 1NA, UK. Tel: 01494 522141 Extension 2349; Fax: 01494 603179; Email: colin.martin@bucks.ac.uk

Measurement equivalence of the Birth Satisfaction Scale-Revised (BSS-R): further evidence of construct validity

Abstract

Objective and background: The 10-item Birth Satisfaction Scale-Revised (BSS-R) is being increasingly used internationally, including the development of translated versions of the tool. However, to date, a direct comparison between the original version of the tool and a non-English language translated version has yet to be conducted. Recognising that measurement equivalence is critical in order to be able to meaningfully compare scores on the measure between different versions, the current sought to evaluate the measurement invariance characteristics of the BSS-R within this context.

Methods: A secondary analysis of two datasets. The study used a measurement invariance testing approach to determine the relative equivalence between the original UK English-language version (Hollins Martin and Martin, 2014) and the Greek-translated version (Vardavaki et al., 2015) of the BSS-R. Participants were a convenience sample of UK (n=228) and Greek (n=162) postnatal women (n=162).

Results: The BSS-R was found to offer an excellent model fit with pooled data, a robust configural model and metric level invariance between English and Greek-language versions. The BSS-R was also found to demonstrate partial scalar invariance, with 80% of item intercepts non-invariant between both versions. Two non-invariant items at the scalar level are likely to represent real differences between participant groups in terms of birth satisfaction as an artefact of service delivery type and relative difference in delivery mode.

Conclusion: The BSS-R is both conceptually and statistically comparable between different versions of the tool suggesting the utility of the measure for International comparative studies.

Key Words: Birth Satisfaction Scale-Revised (BSS), childbearing women, Greek, translation, measurement invariance, measurement equivalence.

Measurement equivalence of the Birth Satisfaction Scale-Revised (BSS-R): further evidence of construct validity

Introduction

Birth satisfaction is a complex construct defining the woman's individual psychological and emotional appraisal of her birth experience (Sawyer et al., 2013). A number of measures have been to assess birth satisfaction, for example, the Mackey Childbirth Satisfaction Rating Scale (MCSRS; Goodman, MacKey and Tavakoli, 2004), the Labor Agentry Scale (LAS; Hodnett & Simmons-Tropea 1987), Six Simple Questions (SSQ; Harvey et al., 2002) and the Perceptions of Care Adjective Check List – Revised (PCACL-R; Redshaw and Martin, 2009). The SSQ and PCACL-R have been recommended as brief measures of birth satisfaction based on reliability and validity characteristics though a caveat to this recommendation is that the tool provides just a single summary overall score of global birth satisfaction (Sawyer et al., 2013). The item set of the PCACL-R has also been criticised for item ambiguity (Siassokos et al., 2009). Examination of the items of the LAS suggest it is conceptually a locus of control scale in contrast to a satisfaction scale. Existing measures have been criticised for lack of coherence to established theoretical models of satisfaction and/or adequate information regarding questionnaire construction and validation (Sawyer et al., 2013).

Recognising the requirement for a theoretically-anchored self-report measure of the birth satisfaction construct for use in clinical practice and applied research, Hollins Martin and Fleming (2011) developed the 30-item birth satisfaction scale (BSS) from an extensive thematic review of the literature, with birth satisfaction perceptions

embedded in the research literature being transcribed into thematically assigned statements and formatted into declarations that could be responded to by childbearing women using a Likert-type format.

Following further work on the measure a short-form version comprising 10-items, the birth satisfaction scale – revised (BSS-R) was developed through an extensive psychometric evaluation of the long-form version of the scale (Hollins Martin and Martin, 2014). Despite its brevity, the BSS-R remains faithful to the original version of the scale, assessing three domains of (i) stress experienced during childbearing, (ii) women's attributes and, (iii) quality of care with a high degree of construct validity and reliability (Hollins Martin and Martin, 2014) and in defined within a woman-centred context.

The increasing awareness of the importance of birth satisfaction to clinical outcome and the provision of short, valid, reliable and theory-bound measure of the concept in the form of the BSS-R has facilitated International interest in this measure. Recent developments of the BSS-R include the validation of a Greek-language translation (Vardavaki et al., 2015) and an English-language version of the tool adapted for use in the United States (Barbosa-Leiker et al., 2015). The findings from these International studies are consistent with those of the instrument development study of Hollins Martin and Martin (2014) indicating the validity and reliability of the BSS-R defined within these countries respective cultures, health economies and maternity services. However, a direct comparison between countries on the measurement aspects of the BSS-R has yet to be conducted, thus the absolute comparability of measure remains unknown. However, given the International interest in the tool and

the increasing likelihood of International comparative studies that may prefer to use this tool, an evaluation of the measurement comparability of the BSS-R across cultures/languages at a fundamentally more psychometrically robust level than has previously been conducted is highly desirable and represents a natural statistical evolution in the development and evaluation of the tool.

Measurement invariance refers to the ability of an instrument, such as a self-report questionnaire to measure the same construct across different groups or within the same group across different observation points (Byrne, 2010). The observation of measurement invariance between groups offers confidence in the application of the construct across groups and moreover, that scores that have been derived from a measure of the construct may directly be compared (Brown, 2015). Measurement invariance thus offers a powerful approach to drawing conclusions from comparison between groups if the instrument is indeed, confirmed to be invariant across groups. Conversely, an instrument that lacks measurement invariance suggests a critical problem in directly comparing both the construct and the scores from different groups (Brown, 2015; Byrne, 2010). Structural equation modelling (SEM) offers a robust and established methodological framework within which to evaluate measurement invariance. The approach commonly taken within the SEM framework is multiple-groups confirmatory factor analysis (MGCFA).

The current study sought to determine the measurement invariance characteristics of the BSS-R across cultures and languages by a secondary analysis of datasets from the original BSS-R development and validation study (Hollins Martin and Martin, 2014) and the Greek-translated version of the instrument (Vardavaki et al., 2015).

The Greek version of the instrument was chosen for comparison since it may be anticipated that the translation process into a non-English language may have a more significant impact on conceptual equivalence than, for example, a comparison with an alternative English-language version of the tool, such as that used in the US (Barbosa-Leiker et al., 2015). The translation process used to develop the Greek version of the BSS/BSS-R included two cycles of translation, the second cycle including minor amendment following feedback gained from a pilot study to evaluate and ensure understanding, comprehension, comparability and equivalence (Vardavaki et al., 2015).

The following research questions were addressed:

(1) Does the BSS-R demonstrate measurement invariance between UK and Greek-language versions of the tool?
(2) Is the BSS-R conceptually equivalent between UK and Greek-language versions?
(3) Can sub-scale and total scores on the BSS-R be directly compared between UK and Greek-language versions?

Method

A secondary analysis of data from the UK (Hollins Martin and Martin, 2014) and Greek-language (Vardavaki et al., 2015) validation studies of the BSS-R.

Participants

The characteristics of the participants (UK study N = 228, Greek study N = 162) from both studies are described in depth in Hollins Martin and Martin (2014) and

Vardavaki et al. (2015) and readers are referred to those papers for related information including study design, clinical characteristics and ethical approvals.

Statistical analysis

Evaluation of measurement invariance requires the testing of increasingly restrictive versions of the underlying measurement model (Brown, 2015; Byrne, 2010). A precursor to this process is to ensure that the underlying measurement model is plausible (Brown, 2010). Since the measurement model for the UK and Greek versions of the BSS-R have been established to be plausible in the published accounts (Hollins Martin and Martin, 2014; Vardavaki et al., 2015), a sequential process of evaluating this established model can be considered. There remains some debate over the use of an initial omnibus model free of constraints between groups (Steenkamp and Baumgartner, 1998) prior to proceeding to increasingly restricted models. The approach however taken in the current study is to first evaluate an omnibus baseline model of pooled UK and Greek BSS-R data to confirm a baseline model that offers acceptable model fit. This approach is taken due to the confirmed measurement models of the BSS-R have been established in separate studies. Following confirmation of an acceptable overall model, a configural invariance model will be evaluated to determine if the factor model and pattern of loadings is equivalent across group. Establishing configural invariance is critical prior to testing a more restrictive metric invariance model, where item-factor loadings are restricted to be the same across groups. Metric invariance has been emphasised to be a necessary condition for demonstrating empirically that the underlying constructs defined by the measurement model have the same intrinsic meaning between groups (Kline, 2005). Following the determination of metric invariance, scalar invariance is then evaluated in which the item intercepts are restricted to be equal across groups.

Evidence of non-invariance in an item at the scalar level is indicative of group differences on the mean of that item even within the context of having comparable values on the factor related to the item itself. Evidence of a good fit to the configural model, metric and measurement invariance establishes strong invariance across groups. It is possible that some items will be invariant across groups while others won't be and this situation is described as partial invariance (Byrne, 2010) contextually defined by the level of invariance testing at which a non-invariant item is identified. Identification of non-invariant items is facilitated by the comparison of the model with the previous (invariant) model (Byrne, 2010). A significant difference in fit between models would indicate evidence of a non-invariant item. Following identification of a non-invariant items modification indices would be scrutinised to elicit the problematic item and the model run again with the restriction between groups for the non-invariant item freed between groups (Byrne, 2010). Should a significant difference between models still be evident, modification indices would be examined again to determine any additional non-invariant item. This process is then repeated until the difference between this model and the previous model which was invariant is non-significant (Hirschfeld and von Brachel, 2014). A challenge within the measurement invariance literature is the criteria for establishing a significant/non-significant difference between models. A Chi-square difference test has often been used to test for model equivalence (non-significant chi-square indicating equivalence), however, it is accepted that chi-square is inflated by sample size (Bollen, 1989). A more contemporary approach is to use the comparative fit index

(CFI, Bentler, 1990) to compare models, with values of ≤ 0.01 indicating measurement invariance (Cheung and Rensvold, 1999).

Fundamental to measurement invariance evaluation is the fit criteria for the underlying models, thus a poorly fitting model would represent a poor model irrespective of measurement invariance characteristics. Multiple goodness of fit tests (Bentler & Bonett, 1980) were used to evaluate the models including the CFI (Bentler, 1990), the root mean squared error of approximation (RMSEA) and the standardised root mean residual (SRMR). CFI values > 0.90 indicates an acceptable fit to the data (Hu & Bentler, 1995), while a CFI ≥ 0.95 indicates a good fit to the data (Hu & Bentler, 1999). RMSEA values ≤ 0.08 indicate acceptable model fit (Browne & Cudeck, 1993), with values of ≤0.05 indicative of good fit (Schumaker & Lomax, 2010). SRMR values ≤ 0.08 indicate acceptable model fit (Hu & Bentler, 1999). Given the influence of sample size and data characteristics on the χ^2 statistic (Hu & Bentler, 1995), model fit determination is generally estimated with reference to CFI and RMSEA (Byrne, 2010; Hooper et al., 2008). A maximum-likelihood (ML) model estimation approach was adopted (Byrne, 2010; Kline, 1993; 2000) since both Hollins Martin and Martin (2014) and Vardavaki et al. (2015) reported generally normally distributed data characteristics in their studies. Statistical analysis was conducted using the R programming language (R Core Team, 2013).

Results

The findings of the measurement invariance testing are summarised in Table 1. The overall model (all data Model 1.) was found to offer an excellent fit to the data. The configural model fit (Model 2.) was found to be acceptable under all model fit criteria. Evaluation of Model 3. against Model 2. revealed no significant difference (ΔCFI ≤1)

thus confirming metric invariance. Model 4. (scalar invariance) was revealed to variant compared to model 3 (ΔCFI >1). Modification indices suggested BSS-R item-3 *'The delivery room staff encouraged me to make decisions about how I wanted my birth to progress'* to be variant thus constraints were relaxed for this item to be freely estimated between groups (Model 5.) Comparison with model 3. revealed model 5. to still be variant thus modification indices were further inspected which suggested BSS-R item-1. *'I came through childbirth virtually unscathed'* to be variant, thus constraints were relaxed on this item also (Model 6.). Comparison of model 6. with model 3. revealed no significant difference between models (ΔCFI ≤1). Appraisal of $\Delta\chi^2$ model comparison findings were consistent with the ΔCFI approach to measurement invariance model determination. Finally, a hierarchical model postulated by Martin and Hollins (2014) with a higher order factor of experience of childbearing was evaluated as an adaption of model 6 (including the relaxed parameter constraints for BSS-R items 3. and 1. This hierarchical model was found to be similar to model 6. in terms of fit characteristics, $\chi^2_{(df = 75)} = 142.49$, CFI = 0.934, RMSEA = 0.068 and SRMR = 0.068. Comparison of BSS-R items 3 and 1 between UK and Greek datasets revealed a significantly higher UK BSS-R item 3 score, $t_{(388)} = 7.35$ (mean difference = 0.77) and a significantly lower UK BSS-R item 1 score, $t_{(388)} = 4.10$ (mean difference = 0.46).

TABLE 1. ABOUT HERE

Discussion

The findings from this secondary analysis offers compelling additional evidence for the transferability of the domains and structure of the original UK English-language version of the BSS-R to the Greek-language version by the appraisal of

measurement equivalence characteristics between the two datasets. Comparison of the measurement invariance characteristics of the two datasets revealed support for partial scalar invariance. The observation of good fit for the configural model and metric invariance between groups indicates that the conceptual meaning of the BSS-R domains as well as the hypothesised tri-dimensional structure of the measure is equivalent between the two versions (Kline, 2005; Vandenberg and Lance, 2000). Moreover, the observation of a more stringent model supporting partial scalar invariance (80% of items invariant at scalar model level), is highly indicative of interpretation of differences/similarities between these two groups in terms of mean scores is fundamentally representative of true differences/similarities in the latent traits being assessed (Brown, 2015; Byrne et al., 1989) rather than measurement error associated with the translation of the instrument or a hitherto unknown characteristic of the participant population recruited for the Greek translation.

A remaining question concerns the two non-invariant items BSS-R item 3. *'The delivery room staff encouraged me to make decisions about how I wanted my birth to progress'* and BSS-R item 1. *'I came through childbirth virtually unscathed'*. Since partial invariance is required to allow scores to be meaningfully compared, these two non-invariant items are of interest in relation to potential mechanisms of invariance in contrast to being a source of measurement confound. Millsap (1998) considered the impact of non-invariance on item intercepts in depth and drew the conclusion that non-invariance of items at the intercept level may be indicative of true score differences rather than intrinsic and systematic measurement error. Looking specifically at these two items in relation to the original datasets, it could be conjectured that the greater degree of birth satisfaction evidenced by BSS-R item 3.

may reflect aspects of the UK health economy that have attempted to place women at the centre of the birth experience for over 30 years since the landmark Changing Childbirth policy document (1993). Conversely, scrutiny of the summary tables from Hollins Martin and Martin (2014) and Vardavaki et al. (2015) reveals a higher comparative percentage of normal vaginal deliveries in the Greek study, thus this may influence responses on BSS-R item 1. due to the lower rate of interventions such as forceps, ventouse or Ceasarian section, thus resulting in a higher mean score on this item. A limitation of the current investigation concerns sample size. Since the investigation utilised two existing datasets a sample size calculation was not performed. It is acknowledged that the sample sizes were modest for SEM approaches to data analysis (Kline, 2000) and ideally, primary studies should consider sample size estimations for SEM at the study design stage. A further limitation was that just two versions of the scale were evaluated. It is credible to consider further evaluation of the invariance characteristics of the BSS-R across versions as further translation and validation studies are conducted, for example, the Turkish language version of the BSS has recently been published revealing excellent validity (Cetin, Sezer and Merih, 2015). However, invariance evaluation would be required to determine the equivalence of this new version of the instrument and the original UK, and indeed other validated versions if meaningful comparisons were to be made.

Conclusion

This secondary analysis represents the first study to investigate empirically, the equivalence of the alternate versions of BSS-R through a robust measurement invariance approach. Given the importance of the birth satisfaction and the

increasing use of the BSS-R Internationally, the current findings foster added confidence in the conceptual and measurement utility of this brief measure and the potential value of the tool for comparative studies.

Acknowledgements

We would like to thank the two anonymous reviewers for their expert insights and extremely helpful advice, comments and suggestions on an earlier version of this manuscript.

References

Barbosa-Leiker, C., Fleming, S., Hollins Martin, C. J., & Martin, C. R. (2015). Psychometric properties of the Birth Satisfaction Scale-Revised (BSS-R) for US mothers. *Journal of Reproductive and Infant Psychology, 33*(5), 504-511.

Belanger-Levesque, M. N., Pasquier, M., Roy-Matton, N., Blouin, S., & Pasquier, J. C. (2014). Maternal and paternal satisfaction in the delivery room: a cross-sectional comparative study. *BMJ Open, 4*(2), e004013-e004013. doi: 10.1136/bmjopen-2013-004013

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the evaluation of covariance structures. *Psychological Bulletin, 88*, 588-606.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods and Research, 17*, 303-316.

Brown, T. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed.). New York: Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternate ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models*.

Byrne, B. M. (2010). *Structural Equation Modeling with AMOS: Basic Concepts, Applications and Programming* (2nd ed.). New York: Routledge/Taylor and Francis Group.

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

Cetin, F. C., Sezer, A., & Merih, Y. D. (2015). The birth satisfaction scale: Turkish adaptation, validation and reliability study. *Northern Clinics of Istanbul, 2*(2), 142-150.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: a reconceptualization and proposed new method. *Journal of Management, 25*(1), 1-27.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255.

Goodman, P., Mackey, M. C., & Tavakoli, A. S. (2004). Factors related to childbirth satisfaction. *Journal of Advanced Nursing, 46*(2), 212-219.

Harvey, S., Rach, D., Stainton, M., Jarrell, J., & Brant, R. (2002). Evaluation of satisfaction with midwifery care. *Midwifery, 18*, 260-267.

Hirschfeld, G., & von Brachel, R. (2014). Multiple-Group confirmatory factor analysis in R: A tutorial in measurement invariance with continuous and ordinal indicators *Practical Assessment, Research and Evaluation, 19*(7).

Hodnett, E. D., & Simmons-Tropea, D. (1987). The labour agentry scale: psychometric properties of an instrument measuring control during childbirth. *Research in Nursing and Health, 10*, 301-310.

Hollins Martin, C. J., & Fleming, V. (2011). The birth satisfaction scale. *International Journal of Health Care and Quality Assurance, 24*(2), 124-135.

Hollins Martin, C. J., & Martin, C. R. (2014). Development and psychometric properties of the Birth Satisfaction Scale-Revised (BSS-R). *Midwifery, 30*(6), 610-619.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53-60.

Hu, L. T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural Equation Modelling: Concepts, Issues and Applications*. Thousand Oaks, CA: Sage.

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.

Kline, P. (1993). *The Handbook of Psychological Testing*. London: Routledge.

Kline, P. (2000). *A Psychometrics Primer*. London: Free Association Books.

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling* (2nd ed.). New York: Guilford Press.

Redshaw, M., & Martin, C. R. (2009). Validation of a perceptions of care adjective checklist. *Journal of Evaluation and Clinical Practice, 15*(2), 281-288.

Sawyer, A., Ayers, S., Abbott, J., Gyte, G., Rabe, H., & Duley, L. (2013). Measures of satisfaction with care during labour and birth: a comparative review. *BMC Pregnancy Childbirth, 13*, 108. doi: 10.1186/1471-2393-13-108

Schumacker, R. E., & Lomax, R. G. (2010). *A Beginner's Guide to Structural Equation Modelling* (3rd ed.). New York: Routledge/Taylor and Francis Group.

Siassakos, D., Clark, J., Sibanda, T., Attilakos, G., Jefferys, A., Cullen, L., Bisson, D., & Draycott, T. (2009). A simple tool to measure patient perceptions of operative birth. *BJOG: An International Journal of Obstetrics & Gynaecology, 116*(13), 1755-1761.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-90.

R. Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: Foundation for Statistical Computing.

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*(1), 4-70.

Vardavaki, Z., Hollins Martin, C. J., & Martin, C. R. (2015). Construct and content validity of the Greek version of the Birth Satisfaction Scale (G-BSS). *Journal of Reproductive and Infant Psychology, 33*(5), 488-503.

Model	χ^2 (df)	Model comparison	$\Delta\chi^2$	Δdf	*p*	RMSEA	SRMR	CFI	ΔCFI	Invariant
1. Overall	50.39(32)	na	na	na	na	0.038	0.032	0.982	na	na
2. Configural	119.53(64)	na	na	Na	na	0.067	0.062	0.945	na	na
3. Metric	127.10(71)	2	7.57	7	0.37	0.064	0.059	0.945	0.000	Yes
4. Scalar	160.07(78)	3	32.97	7	<0.05	0.073	0.066	0.919	0.026	No
5. Partial scalar BSS-R item 3 (intercepts)	147.53(77)	3	20.43	6	<0.05	0.069	0.064	0.931	0.014	No
6. Partial scalar BSS-R items 3 & 1 (intercepts)	137.12(76)	3	10.02	5	0.07	0.064	0.062	0.940	0.005	Yes

Table 1. Evaluation of measurement invariance of the BSS-R by increasingly constrained models (Δ = difference; na = not applicable; Model comparison = comparison of models by level of constraint, e.g. Model 3 (Metric) compared to configural model (Model 2) = Model comparison 2; Invariant = model comparison equivalence, Yes/No).