Martin, C.R., Jefford, E., Hollins martin, C.J. (2020). Crisis, what crisis? Replicability of the key measurement characteristics of the Australian version of the Birth Satisfaction Scale-Revised (BSS-R). *International Journal of Childbirth*. 10(3):140-150. http://dx.doi.org/10.1891/IJCBIRTH-D-20-00006

Crisis, what crisis? Replicability of the key measurement characteristics of the Australian version of the Birth Satisfaction Scale-Revised (BSS-R)

Colin R. Martin[1]

Elaine Jefford[2]

and

Caroline J. Hollins Martin[3]

[1*]Professor of Perinatal Mental Health, Institute for Clinical and Applied Health Research (ICAHR), University of Hull, UK, HU6 7RX.

[2]Research Lead - Midwifery, School of Health & Human Sciences, Southern Cross University, Coffs Harbour, New South Wales 2450, Australia.

[3]Professor in Maternal Health, School of Nursing, Midwifery and Social Care Edinburgh Napier University, Scotland, UK. EH11 4BN.

***Corresponding author**
Address for correspondence:
Professor Colin R. Martin
Institute for Clinical and Applied Health Research (ICAHR)
Rm 329, Allam Medical Building
University of Hull
Hull, HU6 7RX, UK
Email: C.R.Martin@hull.ac.uk
Word count: 3345

Keywords: Birth satisfaction, birth experience, scales, psychometrics, replicability

Abstract

Background: Behavioural and medical science is currently in the grip of a 'replication crisis', circumscribed by the failure to replicate a large proportion of key studies and a consequential impact on confidence in the veracity of the scientific method. Given the contemporary nature of the debate it is surprising that the psychometric properties of commonly used outcome measures have not been evaluated in this context, despite the obvious potential for the measure characteristics of the measures themselves to be a source of error within a study. The current investigation sought to replicate the original validation study of the Australian version of the 10-item Birth Satisfaction Scale-Revised (A-BSS-R) with respect to key psychometric aspects and the issues of replicability.

Methods: A replication study of all quantitative aspects of Jefford et al. (2018) with an increased sample size. Participants were a purposive sample of Australian postnatal women (n=445).

Results: Most key quantitative aspects of the original validation study were found to be replicable and consistent with Jefford et al. (2018), the A-BSS-R was found to have excellent psychometric properties fundamentally mirroring the measurement characteristics observed previously. However, a small number of instances of non-replicability were found.

Conclusions: The A-BSS-R is a valid and reliable measure of the birth satisfaction. Replicability, at least in part, is influenced by participant group characteristics, statistical power, sample size. More focus is required on the influence of self-report measures themselves on the germane aspects of successful study replication.

Introduction

Birth satisfaction represents a complex multi-dimensional construct (Hollins Martin & Martin, 2014) of significant interest to both the clinical and research communities, due to the relationship of the maternal perception of the birth experience to a range of maternal (Anding, Rohrle, Grieshop, Schucking, & Christiansen, 2016; Dale-Hewitt, Slade, Wright, Cree, & Tully, 2012) and neonatal (McDonald et al., 2012) outcomes. Contemporary evidence suggests that the birth satisfaction construct is of importance not only in the postpartum period, but also has an enduring influence in maternal caregiving (Bell, Andersson, Goding, & Vonderheid, 2018), thus the concept is of relevance within a developmental context (Galbally et al., 2017; Parkes, Sweeting, & Wight, 2016).

One contemporary measure of the concept that has gained traction over recent years, due to a combination of convincing theoretical underpinnings, brevity, ease of administration, participant acceptability and generally very good measurement characteristics is the Birth Satisfaction Scale-Revised (BSS-R) (Hollins Martin & Martin, 2014). The BSS-R measures an underlying tri-dimensional and thematically informed structure of (i) stress experienced during child-bearing (SE sub-scale), (ii) women's attributes (WA sub-scale) and (iii) quality of care (QC sub-scale). The BSS-R has demonstrated not only generally exemplary psychometric qualities (C. R. Martin et al., 2017) but also utility, in that the instrument can be used to describe the birth experience in detail through the scores from the three BSS-R sub-scales or, should the clinical or research context require, a total score can be calculated and used to describe overall birth satisfaction (Hollins Martin & Martin, 2014). Indeed, this flexibility in the contextual application of the measure has recently been

demonstrated in a study which revealed excellent veracity for both approaches (sub-scale or total score) to scoring the BSS-R . Further, the BSS-R has been widely translated and validated from the original UK version, adding to the body of evidence regarding the reliability of the measure and authenticity of the measurement characteristics of the tool to the underlying theoretical constructs which supported the development of the measure (Barbosa-Leiker, Fleming, Hollins Martin, & Martin, 2015; Burduli, Barbosa-Leiker, Fleming, Hollins Martin, & Martin, 2017; Goncu Serhatlioglu, Karahan, Hollins Martin, & Martin, 2018; Romero-Gonzalez et al., 2019; Skvirsky, Taubman-Ben-Ari, Hollins Martin, & Martin, 2019; Vardavaki, Hollins Martin, & Martin, 2015).

However, against this backdrop of measurement validity that characterises contemporary studies of the BSS-R, a more general issue concerns all psychological measures including the BSS-R, this being the *replicability crisis* which not only represents a hot topic in the behavioural and medical sciences more generally (Shrout & Rodgers, 2018), but is also highly controversial, with researchers and theoreticians disagreeing on both the magnitude and relevance of the aforementioned *crisis* (Krauss, 2018; Pashler & Harris, 2012; Stroebe & Strack, 2014). The *replicability crisis (RP)* concerns recent observations, originally regarding key psychological studies and experiments, where a significant proportion could not be replicated (Lilienfeld, 2017). Further, it has been observed even within those studies that could be replicated, the effect size is often significantly diminished (Camerer et al., 2018). A number of possible reasons have been suggested for the RP phenomena, with perhaps the most fundamental being, a failure to *actually* replicate studies (Anderson & Maxwell, 2017; Bardi & Zentner, 2017; Coiera,

Ammenwerth, Georgiou, & Magrabi, 2018; Fanelli, 2018; Lilienfeld, 2017; Stanley & Spence, 2014). This observation is perhaps the most surprising from a behavioural science perspective, particularly given that the assumed strength of quantitative research approaches is generalisability, it is thus surprising that the notion of generalisability is simply assumed rather than evaluated at its most fundamental level by a replication study. This issue is conceptually and operationally critically important within the context of healthcare provision and monitoring given that interventions should be based on the best clinical evidence and specifically the ability to assess such efficacy is entirely contingent on measures of outcome. Thus, if the characteristics of health outcomes measures generally and specifically cannot be replicated under the same clinical and/or interventional conditions, then evaluation of healthcare provision cannot be undertaken with confidence, neither can conclusions drawn be assumed to be robust and convincingly erudite.

Finally, an additional explanation given for the lack of replication studies is that peer-reviewed journals would not be interested in them (G. N. Martin & Clarke, 2017). Taking the second part of the last sentence as a challenge, we sought to conduct a replication study mirroring as closely as possible the validation study of the Australian version of the BSS-R (Jefford, Hollins Martin, & Martin, 2018) replicating the methodology, setting and participant profile.

The current investigation sought to replicate the findings of the validation study originally conducted to develop the Australian version of the BSS-R (A-BSS-R). The primary objectives of the current study are to:

1. Demonstrate the replicability of the tri-dimensional measurement model of the original A-BSS-R to a new Australian BSS-R dataset.
2. Evaluate the equivalence between the measurement and structural model of the A-BSS-R between new data and the original validation data of the Australian version of the BSS-R.
3. Evaluate the congruence between the correlational relationships of BSS-R sub-scales between the current dataset and the findings from the A-BSS-R validation study.
4. Evaluate the internal consistency of the Quality of Care (QC), Women's Attributes (WA), and Stress Experienced during Childbearing (SE) sub-scales and the total A-BSS-R scale and compare with the A-BSS-R validation study.
5. Evaluate the known-groups discriminant validity of the A-BSS-R consistent with the approach taken in the Australian A-BSS-R validation study.
6. Evaluate the divergent validity of the A-BSS-R R consistent with the approach taken in the Australian A-BSS-R validation study.

Method

Given that the study is essentially a replication of Jefford et al. (2018), and for the purposes of brevity, readers are guided to that paper for a full review of participant recruitment and statistical nomenclature, however the essential elements are detailed in the current paper. Consistent with Jefford et al. (2018) women were invited to complete the A-BSS-R within six weeks of birth.

Ethical approval

Ethical approval for the study was granted by Southern Cross Research Ethics Committee, Australia.

Participants

Participants were purposively sampled postnatal women who were currently taking part in the Continuity of Care Experience (CoCE) programme. Registration as a midwife in Australia is contingent on engagement with the CoCE. This process involves the midwifery student following a minimum of ten women during the childbearing period in order to gain a unique insight into the practice of midwifery and being with the childbearing woman in a saturated and in-depth manner. The ethos of the CoCE programme, among a number of aspects of contemporary midwifery practice, is to foster a trusting and partnership relationship with the woman throughout the perinatal period. Having given informed consent, study participants were able to access the online survey and complete the A-BSS-R.

Measures

The BSS-R (Hollins Martin & Martin, 2014) is a multi-dimensional ten-item birth satisfaction self-report measure scored on a five-point Likert type scale with responses ranging from (i.) strongly agree, (ii.) agree, (iii.) neither agree or disagree, (iv.) disagree, (v.) strongly disagree with reverse scoring for a number of items. The *stress experienced during childbearing* and *quality of care* sub-scales each comprise four items. The *women's attributes* sub-scale comprises two items. Higher sub-scale and/or total scores indicate comparatively greater birth satisfaction. The Australian version of the BSS-R (A-BSS-R) has also been found to have excellent psychometric properties (Jefford et al., 2018) and found to be conceptually and measurement equivalent to the original United Kingdom version.

Statistical analysis

Confirmatory factor analysis

Objective 1 was addressed using Confirmatory Factor Analysis (CFA) (Brown, 2015). CFA represents a specific case of structural equation modelling (SEM) whereby a conceptual model can be evaluated for goodness of statistical fit against established threshold criteria across a range of indices (Brown, 2015). One of the primary statistical assumptions underlying CFA and SEM is data normality as the approach is circumscribed by parametric assumptions (Brown, 2015). Consequently, data is carefully evaluated for excessive skew and kurtosis and the identification of multivariate outliers which are removed in order to reduce the possibility of violation of the parametric assumptions that are central to this statistical approach and thus reduce the risk of an erroneous interpretation of the analysis. (P. Kline, 2000). Three CFA models were evaluated, which were the tri-dimensional measurement model of the BSS-R comprising correlated factors of SE, WA and QC specified by Hollins Martin and Martin (Hollins Martin & Martin, 2014), a single factor version of this model (correlations between factors set to 1) and a bifactor model (C. R. Martin et al., 2018). Assuming multivariate normality, a maximum-likelihood estimation approach was taken to model evaluation (Brown, 2015; R. B. Kline, 2011). The acceptability of model fit for each of the three models was determined by a range of fit indices (Bentler & Bonett, 1980), including the comparative fit index (CFI) (Bentler, 1990), the root mean squared error of approximation (RMSEA) (Steiger & Lind, 1980) and the square root mean residual (SRMR) (Hu & Bentler, 1999) using threshold conventions for these measures (>0.90, <0.08 and <0.06 respectively).

Invariance analysis

Invariance analysis, similar to CFA, represents an SEM approach to model evaluation and in many respects can be considered CFA for multiple groups or, as in this instance for objective two, across datasets to determine if the characteristics of the data, in terms of the hypothesised model inherent within, is equivalent across groups. To progress to invariance analysis (objective two), an model fit to the CFA must be determined to be acceptable, thus objective one and two are intrinsically related and objective two is contingent on objective one. Following identification of the optimal CFA model (from objective one) increasingly restrictive versions of the BSS-R measurement model are then to be evaluated across datasets (Brown, 2015). CFA and invariance evaluation use identical model fit indices and thresholds thus CFI, RMSEA and SRMR are again used for model evaluation across datasets. Stagewise progression then follows through increasingly restrictive models (Byrne, Shavelson, & Muthen, 1989; Hirschfeld & von Brachel, 2014; C. R. Martin et al., 2017; Vandenberg & Lance, 2000). After establishing that the tri-dimensional model of the BSS-R offers a good fit to data, measurement invariance evaluation will initially evaluate a combined dataset of the data of Jefford et al. (2018) and the current data to determine a good overall fit to data. The next step is to evaluate a configural invariance model to determine the factor model and pattern of loadings is equivalent across both datasets. Item-factor loadings (metric invariance) will then be evaluated for equivalence between datasets in the event configural invariance is established. Satisfactory metric invariance is a requisite for evaluating the more restrictive scalar invariance whereby item-intercepts are evaluated for equivalence between datasets. In the event scalar invariance is established between datasets, strict invariance will be evaluated whereby item-residuals are specified within the model as equivalent

between datasets. Full metric invariance is required to engender confidence in conceptual equivalence of meaning and measurement between data and the transferability of the conceptual and measurement assumptions to another applied context for use of the instrument (Vandenberg & Lance, 2000). Evidence of non-invariance between increasingly restrictive models and associated items is determined by a difference in CFI of >0.01 (Cheung & Rensvold, 2002).

Internal consistency

Cronbach's alpha (Cronbach, 1951) was used to evaluate the internal consistency of the A-BSS-R sub-scales and total score with a threshold of 0.70 or greater acknowledged as being indicative of acceptable internal consistency. Consistent with Jefford et al. (2018) and recognising that alpha is deflated when a scale contains few items (Cortina, 1993; Schmitt, 1996), the two-item WA sub-scale was also evaluated using the inter-item correlation (Pearson's *r*) with reference to Clark and Watson (1995) recommendation that inter-item correlations between 0.15-0.50 indicate scale acceptability. Eisinga, Grotenhuis, and Pelzer (2013) have indicated that this approach is potentially preferable to using Cronbach's alpha in the instance of a two-item scale and was also the approach undertaken in the Australian validation of the BSS-R (Jefford et al., 2018).

Known-groups discriminant validity

Several studies (Fleming et al., 2016; Romero-Gonzalez et al., 2019; Skvirsky et al., 2019; Vardavaki et al., 2015) that have evaluated the known-groups discriminant validity (KGDV) of the BSS-R have examined differences between BSS-R sub-scale and total scale scores as a function of delivery type, dichotomously split between unassisted vaginal delivery (UVD) and intervention delivery (ID; elective Caesarean section (CS), emergency CS, suction cap and instrument) including Jefford et al. (2018). Consistent with the findings of Jefford et al. (2018) it was predicted that the total BSS-R score, and SE and WA sub-scale scores would be significantly higher in the UVD group compared to the ID group while no statistically significant differences between groups is predicted for the QC sub-scale score.

Divergent validity

Divergent validity was evaluated by correlating A-BSS-R total and sub-scale scores with the number of weeks gestation. No statistically significant correlation is predicted with the exception of the WA sub-scale where a statistically significant positive correlation is predicted between this BSS-R sub-scale and gestation as observed in Jefford et al. (2018).

Results

Participants

Five-hundred and twenty-eight eligible participants consented to take part in the study of which N=459 provided complete A-BSS-R data. Screening for multivariate outliers by reference to Mahalanobis distances revealed N=14 outliers which were excluded from the dataset and thus a final dataset of N=445 A-BSS-R complete and multivariate normal cases were prepared for analysis. No statistically significant difference was observed in the relative number of outliers between Jefford et al. (2018) and the current study (χ^2 = 0.56, df = 1, p = 0.46). The mean age of participants was 30.00 (SD 4.79) years with a range of 18-46. The mean duration of pregnancy was 39.34 (SD 1.70) weeks. One-hundred and seventy-nine (40%) women were having their first baby. The means and distributional characteristics of individual A-BSS-R items are summarised in Table 1. A-BSS-R sub-scale and total scale scores are also summarised in Table 1. No evidence of excessive skew or kurtosis was observed.

TABLE 1. ABOUT HERE

Confirmatory factor analysis

The findings of the CFA's reveal an excellent fit to data of both the three-factor measurement model of the BSS-R (χ^2 (df) = 73.87 (32), RMSEA = 0.054, SRMR = 0.043, CFI = 0.976) and the bi-factor model (χ^2 (df) = 64.34 (25), RMSEA = 0.059, SRMR = 0.038, CFI = 0.977). No statistically significant difference in model fit was observed between the three-factor and bi-factor models when using the chi-square differences test ($\Delta\chi^2$ (df) = 9.53 (7), p = 0.22), or the CFI differences approach (ΔCFI = 0.001). A single-factor model was also evaluated and offered a poor fit to the data (χ^2 (df) = 477.94 (35), RMSEA = 0.169, SRMR = 0.106, CFI = 0.744).

Invariance analysis

The findings of the invariance analysis of the A-BSS-R between the current dataset and that of Jefford et al. (2018) are summarised in Table 2. These observations reveal a pattern of generally excellent fit to data within the measurement part of the model from the configural model through to the strict invariance model. Given these findings of comprehensive measurement invariance a post-hoc evaluation of the structural model was undertaken to determine the invariance status of the latent means, latent variances and latent covariances. These also revealed findings of comprehensive structural invariance of the tri-dimensional model of the BSS-R.

TABLE 2. ABOUT HERE

Correlational congruence

Correlations between A-BSS-R sub-scale and total scores in all combinations are summarised in Table 3. and were all positive and statistically significant (p<0.05). Comparisons between the current observations and those reported by Jefford et al. (2018) using the approach of Diedenhofen and Musch (2015) revealed no statistically significant differences between studies on any correlational combination.

TABLE 3. ABOUT HERE

Internal consistency

Cronbach alphas for the A-BSS-R total scale and all sub-scales were at or greater than threshold (0.70). Comparisons between the current study and those of Jefford et al. (2018) are summarised in Table 4. and reveal no statistically significant differences between studies. Inter-item correlation of the A-BSS-R WA sub-scale items was r = 0.56, p <0.001.

TABLE 4. ABOUT HERE

Known-groups discriminant validity

Participants who had an unassisted vaginal delivery were observed to have significantly higher scores on all sub-scales and the total A-BSS-R measure compared to those who had an intervention delivery (Table 5.).

TABLE 5. ABOUT HERE

Divergent validity

No significant correlations were found between the A-BSS-R total and sub-scale scores and the number of weeks gestation, (total) r = -0.01, p = 0.79, (SE) r = -0.06, p = 0.23, (WA) r = 0.03, p = 0.61 and (QC) r = 0.04, p = 0.45.

Discussion

The current investigation offers a number of valuable insights within the context of the contemporary debate regarding replicability and the replicability crisis while providing affirmation of previous psychometric observations regarding the Australian version of the BSS-R (Jefford et al., 2018). Consistent with Jefford et al. (2018) it was found that the tri-dimensional measurement model of the BSS-R offered an excellent fit to the data. Further, no difference was observed between the fit of the tri-dimensional measurement model and the bi-factor model, offering additional support for the position of Martin and colleagues (2018) that the BSS-R total score or sub-scale scores (or indeed both) can be used with equal utility depending on the context of clinical or research application. Confidence in the absolute comparability of both studies is also conferred by the observation of full measurement and structural invariance between the current data set and that of Jefford et al. (2018). Moreover, these psychometric observations represent a critical reflection on replicability, essentially, if parameters between studies are held constant replicability can not only be observed but would be *anticipated with confidence* to be observed. Thus, this finding highlights that a significant potential contributor to the failure to replicate previous study findings is a confound due to differences in population parameters. Krauss (2018) noted that within the context of RCT's that background characteristics of participants are often not adequately distributed between groups leading to a consequential and uncontrolled deleterious impact on outcomes and by implication, replicability.

It was observed in the current observation that not only were correlations between all A-BSS-R sub-scales (and total score) were positively and significantly correlated, but also that no statistically significant differences were observed between the current study *r* values and those reported by Jefford et al. (2018). This contrasts with translation studies such as Romero-Gonzalez and colleagues (2019) where significant differences between the correlational relationship of the Spanish total BSS-R score to the BSS-R SE sub-scale were observed in comparison to the original UK development study (Hollins Martin & Martin, 2014), thus indicating the potential influence of population characteristics in replicability.

Internal consistency observations in the current study were all observed to be acceptable and moreover no statistically significant differences were observed between these internal consistency observations and those reported by Jefford et al. (2018). Contrasting again with Romero-Gonzalez et al. (2019) where significant differences were observed between Cronbach's alpha between the Spanish BSS-R QC sub-scale and the original English-language version (Hollins Martin & Martin, 2014). Noting that in Jefford and colleagues A-BSS-R development study no statistically significant differences were observed between the above correlational relationships and internal consistency estimates and the original UK scale development study (2014), it may be inferred that cultural context and translational process may also play a subtle role in replicability issues particularly if the sample and/or translational process is not adequately described even when the psychometric profile of the instrument is exemplary in that group. Further, one advantage of the BSS-R is its brevity, being 10-items, however this also means that changes to just one item in the translational process represents a 10% change in wording beyond purely literal translation. It is not unusual in BSS-R translation/validation studies for two items to be changed (Barbosa-Leiker et al., 2015; Jefford et al., 2018; Nespoli, Colciago, Pedroni, Perego, & Fumagalli, 2018) thus representing a 20% change to the measure and we would suggest at the very least, a consideration should be taken of any impact on replicability in a study at the level of the instrument itself, a point of particular importance when tools such as the BSS-R are used as outcome measures within the milieu of International comparisons (Nijagal et al., 2018).

A pertinent contribution to the debate on issues of replicability has been highlighted in relation to effect sizes and adequate statistical power. Anderson and Maxwell (2017) have suggested that sample size calculations for replication studies are often unduly optimistic due to factors such as publication bias and uncertainty when such calculations are conducted from the original study. The findings from the current study offer a useful insight on this position in terms of the known-group discriminant validity observations. It was observed that those who had a UVD had significantly higher A-BSS-R scores than those that had an ID across all sub-scales and the total score. This contrasts with Jefford et al. (2018) where comparisons between groups revealed A-BSS-R QC sub-score differences to be non-significant between groups (p = 0.07). Further, though the effect size is small in relation to the A-BSS-R QC sub-scale comparison in this study and that of Jefford et al. (2018), it is likely that the current study large sample size offers sufficient statistical power to detect statistical significance even in relation to a small effect. Thus in relation to the A-BSS-R QC sub-scale we were not able to replicate the findings of Jefford et al. (2018) even though descriptively group mean scores between studies were almost identical. This observation not only vindicates the position of Anderson and Maxwell (2017) and indeed others (Grabitz et al., 2018; Shrout & Rodgers, 2018) regarding appropriate sample size but also highlights the need to consider the limitations of current approaches to sample size estimation and the desirability to consider novel methodologies to address this particular issue (Davis-Stober & Dana, 2014).

Divergent validity analysis found no significant relationship between A-BSS-R total and sub-scale scores and number of weeks gestation. One inconsistency was thus found with the observations of Jefford et al. (2018) who observed a significant relationship in relation to the A-BSS-R WA sub-scale ($r = 0.17$, $p = 0.01$). Considering that the relationship between effect size, power, alpha and sample size is invariably linked to postulating differences between groups/variables, the notion that increased sample size may be associated with effect size deflation has received little attention.

A reflection on a fundamental aspect of the replication focus of the current investigation is that key variables, for example age and duration of pregnancy were essentially identical between this study and that of Jefford et al. (2018). It is noteworthy that in other studies that have used the BSS-R, for example (Skodova, Nepelova, Grendar, & Baskova, 2019) and (Hamm, Srinivas, & Levine, 2019), the age of participants has been similar to those in the current study, whereas there have been a number of BSS-R studies where participants have been somewhat younger, for example, Barbosa-Leiker et al. (2015) and Goncu Serhatlioglu et al. (2018) or older (Fleming et al., 2016; Romero-Gonzalez et al., 2019). However, with regard to variation in this key variable it is important to acknowledge that within the context of psychometric studies such variability would appear to have little impact on the fundamental measurement characteristics of the BSS-R, particularly in terms of the tri-dimensional measurement model and accompanying fit to data (Goncu Serhatlioglu et al., 2018; Romero-Gonzalez et al., 2019). In terms of duration of pregnancy, published studies that have used the BSS-R have generally reported a similar gestational timeline to those reported herein and by Jefford et al. (2018) and therefore the potential impact of duration of pregnancy on birth satisfaction may perhaps only be illuminated in terms of significant impact by examination of specific clinical groups of interest, for example, those women experiencing premature birth.

Finally, it is noted that the purposive sampling approach (participants recruited exclusively through the CoCE programme) represents a limitation within the study in terms of confidence in the extent of generalisability compared to probabilistic or random sampling. However, given the inherent foci of the study was on replication of Jefford et al. (2018) study which utilised the same purposive sampling approach, this limitation also paradoxically represents a strength of the current investigation.

Conclusion

This investigation represents the first study (as far as we are aware) to investigate the contemporary issue of the replication crisis within the context of maternity care and measurement of birth satisfaction. Through following identical data collection procedures and parameters as the original study (Jefford et al., 2018), the investigation replicated the majority of the psychometric findings of that study. However, there were some differences noted, even following as stated, an identical research protocol. The most parsimonious explanation for the differences between studies observed in known-group discriminant validity analysis is the larger sample size of the study thus increasing the power to detect a significant difference. More theoretically vexing however, was a difference between studies on one assessment of divergent validity, an area which may benefit from both theoretical insight and statistical innovation. Largely and taken in the round, our observations support the methodological foundation stone of quantitative research that replication is achievable when strict adherence to study design parameters is exercised. Further research is clearly desirable to examine more closely the replicability of the psychometric properties of commonly used self-report measures between studies including the effect of background factors in influencing findings.

Obtaining the A-BSS-R

To request a copy of the A-BSS-R and the associated A-BSS-R scoring grid, please contact Professor Caroline Hollins Martin by email: C.HollinsMartin@napier.ac.uk

References

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "Replication Crisis": Using Original Studies to Design Replication Studies with Appropriate Statistical Power. *Multivariate Behavioral Research, 52*(3), 305-324. doi:10.1080/00273171.2017.1289361

Anding, J. E., Rohrle, B., Grieshop, M., Schucking, B., & Christiansen, H. (2016). Couple comorbidity and correlates of postnatal depressive symptoms in mothers and fathers in the first two weeks following delivery. *Journal of Affective Disorders, 190*, 300-309. doi:10.1016/j.jad.2015.10.033

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. doi:10.1080/02646838.2015.1024211

Bardi, A., & Zentner, M. (2017). Grand Challenges for Personality and Social Psychology: Moving beyond the Replication Crisis. *Frontiers in Psychology, 8*, 2068. doi:10.3389/fpsyg.2017.02068

Bell, A. F., Andersson, E., Goding, K., & Vonderheid, S. C. (2018). The birth experience and maternal caregiving attitudes and behavior: A systematic review. *Sexual and Reproductive Healthcare, 16*, 67-77. doi:10.1016/j.srhc.2018.02.007

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2320703

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

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

Burduli, E., Barbosa-Leiker, C., Fleming, S., Hollins Martin, C. J., & Martin, C. R. (2017). Cross-cultural invariance of the Birth Satisfaction Scale-Revised (BSS-R): comparing UK and US samples. *Journal of Reproductive and Infant Psychology, 35*(3), 248-260. doi:10.1080/02646838.2017.1310374

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.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*. doi:doi.org/10.1038/s41562-018-0399-z

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. doi:10.1207/S15328007SEM0902_5

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development *Psychological Assessment, 7*(3), 309-319.

Coiera, E., Ammenwerth, E., Georgiou, A., & Magrabi, F. (2018). Does health informatics have a replication crisis? *Journal of the American Medical Informatics Association*. doi:10.1093/jamia/ocy028

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 79*, 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Dale-Hewitt, V., Slade, P., Wright, I., Cree, M., & Tully, C. (2012). Patterns of attention and experiences of post-traumatic stress symptoms following childbirth: an experimental study. *Archives of Womens Mental Health, 15*(4), 289-296. doi:10.1007/s00737-012-0290-2

Davis-Stober, C. P., & Dana, J. (2014). Comparing the accuracy of experimental estimates to guessing: a new perspective on replication and the "Crisis of Confidence" in psychology. *Behavior Research Methods, 46*(1), 1-14. doi:10.3758/s13428-013-0342-1

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS One, 10*(4), e0121945. doi: doi:10.1371/journal.pone.0121945

Eisinga, R., Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health, 58*(4), 637-642. doi:10.1007/s00038-012-0416-3

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences of the United States of America, 115*(11), 2628-2631. doi:10.1073/pnas.1708272114

Fleming, S. E., Donovan-Batson, C., Burduli, E., Barbosa-Leiker, C., Hollins Martin, C. J., & Martin, C. R. (2016). Birth Satisfaction Scale/Birth Satisfaction Scale-Revised (BSS/BSS-R): A large scale United States planned home birth and birth centre survey. *Midwifery, 41*, 9-15. doi:10.1016/j.midw.2016.07.008

Galbally, M., van, I. M., Permezel, M., Saffery, R., Lappas, M., Ryan, J., . . . Lewis, A. J. (2017). Mercy Pregnancy and Emotional Well-being Study (MPEWS): Understanding maternal mental health, fetal programming and child development. Study design and cohort profile. *International Journal of Methods in Psychiatric Research, 26*(4). doi:10.1002/mpr.1558

Goncu Serhatlioglu, S., Karahan, N., Hollins Martin, C. J., & Martin, C. R. (2018). Construct and content validity of the Turkish Birth Satisfaction Scale - Revised (T-BSS-R). *Journal of Reproductive and Infant Psychology*, 1-11. doi:10.1080/02646838.2018.1443322

Grabitz, C. R., Button, K. S., Munafo, M. R., Newbury, D. F., Pernet, C. R., Thompson, P. A., & Bishop, D. V. M. (2018). Logical and Methodological Issues Affecting Genetic Studies of Humans Reported in Top Neuroscience Journals. *Journal of Cognitive Neuroscience, 30*(1), 25-41. doi:10.1162/jocn_a_01192

Hamm, R. F., Srinivas, S. K., & Levine, L. D. (2019). Risk factors and racial disparities related to low maternal birth satisfaction with labor induction: a prospective, cohort study. *BMC Pregnancy and Childbirth, 19*(1), 530. doi:10.1186/s12884-019-2658-z

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). Retrieved from http://pareonline.net/getvn.asp?v=19&n=7

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. doi:10.1016/j.midw.2013.10.006

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.

Jefford, E., Hollins Martin, C. J., & Martin, C. R. (2018). Development and validation of the Australian version of the Birth Satisfaction Scale-Revised (BSS-R). *Journal of Reproductive and Infant Psychology, 36*(1), 42-58. doi:10.1080/02646838.2017.1396302

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

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (3rd ed.). London: Guilford Press.

Krauss, A. (2018). Why all randomised controlled trials produce biased results. *Annals of Medicine, 50*(4), 312-322. doi:10.1080/07853890.2018.1453233

Lilienfeld, S. O. (2017). Psychology's Replication Crisis and the Grant Culture: Righting the Ship. *Perspectives on Psychological Science, 12*(4), 660-664. doi:10.1177/1745691616687745

Martin, C. R., Hollins Martin, C. J., Burduli, E., Barbosa-Leiker, C., Donovan-Batson, C., & Fleming, S. E. (2017). Measurement and structural invariance of the US version of the Birth Satisfaction Scale-Revised (BSS-R) in a large sample. *Women and Birth, 30*(4), e172-e178. doi:10.1016/j.wombi.2016.11.006

Martin, C. R., Hollins Martin, C. J., Burduli, E., Barbosa-Leiker, C., Donovan-Batson, C., & Fleming, S. E. (2018). The Birth Satisfaction Scale - Revised (BSS-R): should the subscale scores or the total score be used? *Journal of Reproductive and Infant Psychology*, 1-6. doi:10.1080/02646838.2018.1490498

Martin, G. N., & Clarke, R. M. (2017). Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices. *Frontiers in Psychology, 8*, 523. doi:10.3389/fpsyg.2017.00523

McDonald, S., Wall, J., Forbes, K., Kingston, D., Kehler, H., Vekved, M., & Tough, S. (2012). Development of a prenatal psychosocial screening tool for post-partum depression and anxiety. *Paediatric and Perinatal Epidemiology, 26*(4), 316-327. doi:10.1111/j.1365-3016.2012.01286.x

Nespoli, A., Colciago, E., Pedroni, S., Perego, S., & Fumagalli, S. (2018). The Birth Satisfaction Scale-Revised (BSS-R): process of translation and adaptation in an Italian context. *Annali dell'Istituto Superiore di Sanità, 54*(4), 340-347. doi:10.4415/ANN_18_04_11

Nijagal, M. A., Wissig, S., Stowell, C., Olson, E., Amer-Wahlin, I., Bonsel, G., . . . Franx, A. (2018). Standardized outcome measures for pregnancy and childbirth, an ICHOM proposal. *BMC Health Services Research, 18*(1), 953. doi:10.1186/s12913-018-3732-3

Parkes, A., Sweeting, H., & Wight, D. (2016). What shapes 7-year-olds' subjective well-being? Prospective analysis of early childhood and parenting using the Growing Up in Scotland study. *Social Psychiatry and Psychiatric Epidemiology, 51*(10), 1417-1428. doi:10.1007/s00127-016-1246-z

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science, 7*(6), 531-536. doi:10.1177/1745691612463401

Romero-Gonzalez, B., Peralta-Ramirez, M. I., Caparros-Gonzalez, R. A., Cambil-Ledesma, A., Hollins Martin, C. J., & Martin, C. R. (2019). Spanish validation and factor structure of the Birth Satisfaction Scale-Revised (BSS-R). *Midwifery, 70*, 31-37. doi:10.1016/j.midw.2018.12.009

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology, 69*, 487-510. doi:10.1146/annurev-psych-122216-011845

Skodova, Z., Nepelova, Z., Grendar, M., & Baskova, M. (2019). Psychometric properties of the Slovak version of the Birth Satisfaction Scale (BSS) and Birth Satisfaction Scale-Revised (BSS-R). *Midwifery, 79*, 102550. doi:10.1016/j.midw.2019.102550

Skvirsky, V., Taubman-Ben-Ari, O., Hollins Martin, C. J., & Martin, C. R. (2019). Validation of the Hebrew Birth Satisfaction Scale - Revised (BSS-R) and its relationship to perceived traumatic labour. *Journal of Reproductive and Infant Psychology*, 1-7. doi:10.1080/02646838.2019.1600666

Stanley, D. J., & Spence, J. R. (2014). Expectations for Replications: Are Yours Realistic? *Perspectives on Psychological Science, 9*(3), 305-318. doi:10.1177/1745691614528518

Steiger, J. H., & Lind, J. (1980). *Statistically-based tests for the number of common factors*. Paper presented at the Annual Spring Meeting of the Psychometric Society, Iowa City, USA.

Stroebe, W., & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science, 9*(1), 59-71. doi:10.1177/1745691613514450

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. doi:doi: 10.1177/109442810031002

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. doi:10.1080/02646838.2015.1035235

Table 1. Mean, standard deviation and distributional characteristics of individual A-BSS-R items, sub-scale totals and the total A-BSS-R score. se = standard error of kurtosis.

Item	Item content	Domain*	Mean	SD	Min	Max	Skew	Kurtosis	se
BSS-R 1	I came through childbirth virtually unharmed	SE	3.09	1.11	0	4	-1.10	0.22	0.05
BSS-R 2	I thought my labour was excessively long	SE	2.81	1.17	0	4	-0.82	-0.14	0.06
BSS-R 3	The birthing room staff encouraged me to make decisions about how I wanted my birth to progress	QC	3.32	0.86	0	4	-1.23	1.11	0.04
BSS-R 4	I felt very anxious during my labour and birth	WA	2.53	1.20	0	4	-0.46	-0.78	0.06
BSS-R 5	I felt well supported by staff during my labour and birth	QC	3.69	0.59	1	4	-1.91	3.34	0.03
BSS-R 6	The staff communicated well with me during labour	QC	3.60	0.66	1	4	-1.71	2.75	0.03
BSS-R 7	I found giving birth a distressing experience	SE	2.76	1.13	0	4	-0.67	-0.44	0.05
BSS-R 8	I felt out of control during my birth experience	WA	2.84	1.08	0	4	-0.71	-0.38	0.05
BSS-R 9	I was not distressed at all during labour	SE	2.16	1.16	0	4	0.03	-0.93	0.05
BSS-R 10	The delivery room was clean and hygienic	QC	3.76	0.47	2	4	-1.81	2.41	0.02
Stress	Sub-scale total		10.83	3.52	0	16	-0.52	-0.22	0.17
Attributes	Sub-scale total		5.37	2.01	0	8	-0.50	-0.45	0.10
Quality	Sub-scale total		14.37	2.03	7	16	-1.26	0.93	0.10
Total	Total score		30.58	6.29	11	40	-0.53	-0.19	0.30

*Domain of the A-BSS-R. SE = Stress experienced during child-bearing, WA = Women's attributes, QC = Quality of Care.

Table 2. Invariance analysis of Jefford et al. (2018) and current study BSS-R datasets.

Model	χ^2 (df)	Model comparison	$\Delta\chi^2$	Δdf	*p*	RMSEA	SRMR	CFI	ΔCFI	Different
1. Overall	101.64(32)	na	na	na	na	0.058	0.041	0.971	na	na
2. Configural	145.06(64)	na	na	Na	Na	0.063	0.044	0.967	na	na
3. Metric	158.79(71)	2	13.73	7	0.06	0.062	0.051	0.964	0.003	No
4. Scalar	167.64(78)	3	8.85	7	0.26	0.060	0.052	0.963	0.001	No
5. Strict	210.87(88)	4	43.23	10	0.001	0.066	0.062	0.950	0.013	Yes
6. Partial strict BSS-R 10	192.80(87)	4	25.17	9	0.003	0.062	0.054	0.957	0.006	No

Table 3. Correlations of A-BSS-R sub-scales and total score and comparison with original Australian BSS-R validation study (Jefford et al., 2018).

Scale combination	Jefford et al. *r*	Current study *r*	Z	95% CI	*p*
Stress-Attributes	0.67	0.75	1.89	(-0.01 – 0.17)	0.06
Stress-Quality	0.23	0.36	1.65	(-0.03 – 0.29)	0.10
Attributes-Quality	0.32	0.41	1.21	(-0.05 – 0.24)	0.23
Total score-Stress	0.89	0.91	1.23	(-0.01 – 0.06)	0.22
Total score-Attributes	0.83	0.87	1.69	(-0.01 – 0.09)	0.09
Totals score-Quality	0.60	0.66	1.16	(-0.04 – 0.17)	0.25

Table 4. Cronbach's alpha of BSS-R sub-scales and total score and comparison with original Australian BSS-R validation study (Jefford et al., 2018). Degrees of freedom = 1.

Subscale	Jefford et al. alpha	Current study alpha	χ^2	*p*
Stress	0.74	0.77	0.62	0.43
Attributes	0.66	0.71	0.58	0.45
Quality	0.81	0.76	2.16	0.14
Total score	0.81	0.84	1.68	0.19

Table 5. Comparison of BSS-R total and sub-scale scores differentiated by birth delivery type. Standard deviations are in parentheses, degrees of freedom = 442, CI = confidence interval.

BSS-R Scale	Unassisted vaginal delivery (N=310)	Assisted/ Operative delivery (N=134)	95% CI	*t*	*p*	Hedges g	Hedges g 95% CI	Effect size
Stress	11.75 (3.12)	8.71 (3.49)	2.38 - 3.69	9.07	<0.001	0.93	0.72 - 1.15	Large
Attributes	5.82 (1.81)	4.34 (2.08)	1.11 - 1.87	7.56	<0.001	0.78	0.57 - 0.99	Medium
Quality	14.63 (1.90)	13.75 (2.19)	0.47 - 1.28	4.27	<0.001	0.44	0.24 - 0.65	Small
Total score	32.19 (5.63)	26.80 (6.16)	4.22 - 6.57	9.01	<0.001	0.93	0.72 - 1.14	Large