Establishing a coherent and replicable measurement model of the Edinburgh Postnatal Depression Scale

Colin R. Martin , Maggie Redshaw

PII: S0165-1781(17)32130-3
DOI: 10.1016/j.psychres.2018.03.062
Reference: PSY 11294

To appear in: *Psychiatry Research*

Received date: 20 November 2017
Revised date: 9 March 2018
Accepted date: 22 March 2018

Please cite this article as: Colin R. Martin , Maggie Redshaw , Establishing a coherent and replicable measurement model of the Edinburgh Postnatal Depression Scale, *Psychiatry Research* (2018), doi: 10.1016/j.psychres.2018.03.062

Highlights:

- Standard factor analytic methods were used to explore the measure structure
- Using the same methods in a recent study with two groups of women it was possible to make direct comparisons between responses at key time points after birth
- A three factor model of the EPDS fitted the data best with both data sets
- The three factors identified could be used in screening practice and tested further in research studies

Establishing a coherent and replicable measurement model

of the Edinburgh Postnatal Depression Scale

Colin R. Martin[a]

Maggie Redshaw[b]

[a] Professor of Mental Health, Faculty of Society and Health, Buckinghamshire New University, UK.

Email: colin.martin@bucks.ac.ul

[b] *Senior Research Fellow, Policy Research Unit in Maternal Health and Care, National Perinatal Epidemiology Unit (NPEU), Nuffield Department of Population Health, University of Oxford, UK.

Email: maggie.redshaw@npeu.ox.ac.uk

*Corresponding author

Address for correspondence:

Policy Research Unit in Maternal Health and Care, National Perinatal Epidemiology Unit (NPEU), Nuffield Department of Population Health, University of Oxford, Old Road Campus, Headington, Oxford, UK, OX3 7LF.

ABSTRACT

The 10-item Edinburgh Postnatal Depression Scale (EPDS) is an established screening tool for postnatal depression. Inconsistent findings in factor structure and replication difficulties have limited the scope of development of the measure as a multi-dimensional tool. The current investigation sought to robustly determine the underlying factor structure of the EPDS and the replicability and stability of the most plausible model identified. A between-subjects design was used. EPDS data were collected postpartum from two independent cohorts using identical data capture methods. Datasets were examined with confirmatory factor analysis, model invariance testing and systematic evaluation of relational and internal aspects of the measure. Participants were two samples of postpartum women in England assessed at three months (*n*=245) and six months (*n*=217). The findings showed a three-factor seven-item model of the EPDS offered an excellent fit to the data, and was observed to be replicable in both datasets and invariant as a function of time point of assessment. Some EPDS sub-scale scores were significantly higher at six months. The EPDS is multi-dimensional and a robust measurement model comprises three factors that are replicable. The potential utility of the sub-scale components identified requires further research to identify a role in contemporary screening practice.

Key Words: Edinburgh Postnatal Depression Scale, EPDS, factor structure, validity, psychometrics

Establishing a coherent and replicable measurement model

of the Edinburgh Postnatal Depression Scale

1. Introduction

Postnatal depression (PND) represents a significant mental health concern with an average of 13% of women experiencing this distressing condition O'Hara and Swain (1996), though reported rates differ considerably, for example Banti et al. (2011). The impact of PND is pervasive, with robust evidence of deleterious impact not only on the woman herself (Pope et al, 2013; Wisner et al., 2013), but also on her baby (Dahlen et al., 2015; Fairbrother and Woody, 2008; Jennings et al, 1999; Milgrom and Holt, 2014) and partners (Cameron et al, 2017). Paradoxically, given the impact of PND, universal screening for all postnatal women is currently not policy (American College of Obstetricians and Gynecologists' Committee on Obstetric American College of Obstetricians and Gynecologists' Committee on Obstetric Practice, 2015), current practice in the UK being to consider a brief screen by health professionals using two identification questions and a follow up to a positive response to either question with a validated screening measure or a referral (National Institute for Health and Care Excellence, 2015). The most widely used screening measure for PND is the Edinburgh Postnatal Depression Scale (EPDS) developed by J. L. Cox et al (1987). A driver in the development of the EPDS was the avoidance of items which could be influenced by physical symptoms (J. L. Cox et al., 1987), a critical aspect in approaching screening given the significant physiological changes that accompany pregnancy and childbirth. The EPDS has endured as the most widely used PND screening measure (Moraes et al, 2017; Smith et al, 2016).

Despite, the extensive use of the EPDS as a screening instrument, the measure has also been noted for some contradictory observations in terms of its measurement structure. The measure itself was originally developed to be a unitary measure of (postnatal) depression, however, a multitude of studies have demonstrated the EPDS to have an underlying multi-dimensional factor structure (Brouwers et al, 2001; Gollan et al., 2017; Jomeen and Martin, 2007; Matthey, 2008;

Phillips et al, 2009; Reichenheim et al, 2011; Ross et al, 2003; Tuohy and McVey, 2008). The findings of such studies constructively suggest that the EPDS may comprise sub-scale domains of potential and added clinical value (Matthey, 2008). At the same time they indicate that the tool itself does not appear to measure what it was designed to measure (depression) and consequently may be limited in terms of both screening effectiveness (Matthey and Agostini, 2017) and links to a coherent clinical and unidimensional model of postnatal depression (Gollan et al., 2017). Nevertheless, the notion of a multi-dimensional underlying structure to the EPDS need not necessarily detract from its clinical utility, with identification of robust independent sub-scales embedded within the tool not anticipated by the instrument developers (Matthey, 2008). However, there must be consideration of structural stability, and the multidimensional structure of the EPDS and the embedded sub-scales, should be replicable across groups, for example, depressed/non-depressed, white/black minority ethnic, high social economic status (SES)/low SES (Matthey and Agostini, 2017). This has not been found to be the case, with evidence of wide variation in the items assigned to factors across a range of studies, even within the context of two-factor, or three-factor models which have been the most pervasive factorial determinations of measurement studies of the EPDS (Chabrol and Teissedre, 2004; Jomeen and Martin, 2007; Pallant et al, 2006; Ross et al., 2003; Tuohy and McVey, 2008). Interpretation of the content of underlying factor domains within the EPDS has thus been problematic, due to inconsistent factor structure, with most two factor model solutions reporting domains of anxiety and depression, though the domains themselves have been indicated by different individual items across studies (Reichenheim et al., 2011). Clearly, such unreconciled differences across studies are unsatisfactory in terms of theoretical coherence and practical clinical interpretation. The possibility that the underlying structure of the EPDS may indeed map onto a theoretically robust multi-dimensional model of depression could be inferred by the study of Tuohy and McVey (2008) who described the third factor in their tri-dimensional analysis as representing 'anhedonia'. This observation is not only consistent with an important component of the tri-dimensional model of depression suggested by Clark and Watson (1991) but also resonates with the

finding of a tri-dimensional structure which includes an anhedonia domain to the Hospital Anxiety and Depression Scale (Zigmond and Snaith, 1983), another screening measure that has been frequently used within the perinatal field (Jomeen and Martin, 2008a; Meades and Ayers, 2011; Tohotoa et al., 2012). Reichenheim et al. (2011) conducted an elegant study examining the underlying factor structure of the EPDS, finding evidence for three factors but ultimately recommending the use of a unitary total score to best represent the measurement model of tool. This was premised on the basis of a superior fit of a bi-factor model comprising a general factor and three specific factors, however it has been suggested that superior fit of bifactor models could be due to a 'method effect' in contrast to the empirical superiority of the underlying model which should be specified on conceptual and theoretical grounds (Morgan et al, 2015).

It is noteworthy also that the majority of studies examining the measurement properties of the EPDS have been cross-sectional in design. This is important as the recommendations of not only when to screen for PND but indeed, when PND may be diagnosed as a disorder distinct from major depressive disorder vary dramatically from birth to twelve months depending on the timing of assessment. A number of these cross-sectional studies have recruited across a broad sample range post-partum, for example, women from birth to ten months post-partum (Hartley et al, 2014), between birth and one year (Phillips et al. (2009) and much closer to the birth at 2-3 days postpartum (Teissedre and Chabrol, 2004). For such studies to be compared, a fundamental assumption must be that the underlying structure should be consistent across time. In a large sample (N~1200) study strong evidence was found for a tri-dimensional structure to the EPDS that was consistent in both antenatal and postnatal samples (Coates et al, 2016).

A relatively small number of studies have examined the longitudinal structure of the EPDS and findings from these studies are potentially helpful given the clinical reality of variations in screening times and screening opportunities for PND. A study was conducted on the measurement of women's mental health at admission and at discharge to psychiatric mother and baby units

(Cunningham et al, 2015). Uniquely, this study focused on a clinical group with a confirmed psychiatric diagnosis and incorporating implicitly the effect of intervention on outcome. It was observed that the EPDS comprised two distinct factors on admission and three distinct factors on discharge and concluded that women may interpret EPDS items in characteristically different ways as a function of their degree of psychological/psychiatric distress (Cunningham et al, 2015). The finding from this study that the EPDS measures different constructs at different time points is far-reaching in terms of screening practice and research. However, an important caveat, recognised by the investigators themselves, was that the sample represented a distinct population with diagnosed and significant mental illness requiring in-patient admission and that of course, consequentially, therapeutic intervention represented an inevitable component of the journey between admission and discharge. It is therefore difficult to conclude whether factorial instability would generalise to populations without severe mental illness (Cunningham et al, 2015). This is particularly salient given that the majority of women following birth do not develop PND and the time that they may be screened for PND may vary. A critical issue, specifically, is whether the most robust empirically-derived factorial structure of the EPDS is replicable and consistent in normal population samples drawn at different postpartum intervals.

The objectives of the current study are to:

1. Evaluate comparative model fit of empirically-derived multi-dimensional models of the EPDS against the single factor model.

2. Evaluate comparative model fit of the equivalent tri-dimensional model of the EPDS against a bifactor model of the EPDS as proposed by Reichenheim et al. (2011).

3. Demonstrate the replicability and stability of the best-fit model of the EPDS across time.

4. Determine the measurement coherence of mean EPDS scores across time points.

5. Evaluate the equivalence of EPDS total and sub-scale internal consistency across time points.

6. Determine the equivalence of EPDS total and sub-scale correlational relationships between time points.

7. Evaluate case classification rate concordance between the conventional 10-item EPDS and the recent 7-item EPDS suggested by Gollan et al. (2017).

2. Methods

2. 1 Participants

Data were collected from a randomly selected sample of women in England at either three months (time point 1) or six months (time point 2) postpartum, these being two separate samples thus the use of a between-subjects design. The sample was drawn by the Office for National Statistics who managed the mailing. A questionnaire was sent to each woman selected, with an invitation letter and an information leaflet, followed by a further questionnaire and reminder as appropriate. Women aged less than 16 years were excluded as were those whose babies had died in the months after birth. Completion of the questionnaire was taken as implicit consent to participate. No incentives were offered for questionnaire return.

2.2 Design

A between-subjects design was used to investigate the study objectives in this secondary analysis study. To address objectives 1 (evaluate EPDS model fit) and 2 (tri-dimensional/bifactor model comparison) data were collapsed between time points and single-factor, tri-dimensional and bifactor models compared. To address objective 3 (replicability and stability of best-fit model) the most convincing model found evaluating objectives 1 and 2 would be evaluated using data stratified between the two time points where the stability and replicability of the factor structure would be investigated. Objective 4 (measurement coherence of mean EPDS scores across time points) would be addressed by examining mean EPDS total and sub-scale score differences between the three month and six month observation points. Objective 5 (EPDS total and sub-scale internal reliability)

would be met by comparing internal consistency estimations across the two observation points. Objective 6 (equivalence of EPDS total/sub-scale correlational relationships) would be achieved by comparing correlations between scales and sub-scales across the two observation points. Objective 7 (case classification rate concordance) would be achieved by a comparison of threshold classification rate for the EPDS and EPDS-7.

2.3 Ethical approval

Ethical approval for the survey of infant and maternal health was obtained from the NRES committee for Yorkshire and The Humber – Sheffield Research Ethics Committee (REC reference 16/YH/0412).

2.4 Measures

The EPDS (J. L. Cox et al., 1987) is a 10-item self-report screening tool measure of PND. The cut-off criteria for screening has been noted to vary between studies and clinical populations (Gollan et al., 2017; Matthey and Agostini, 2017), however, the 'classic' cut-point threshold is 12/13for probable PND (J. L. Cox et al., 1987). A 9/10 cut-point for screening has been suggested (J.L. Cox and Holden (2003). Each item is scored on a 4-point Likert type agreement scale (scored 0-3) with some of the items reverse scored. A total score (range 0-30) is calculated on which the cut-point is used to stratify groups (screen negative/screen positive) contingent on the threshold cut-point chosen. All sub-scales and the EPDS-7 are derived from the EPDS.

2.5 Statistical analysis

Objectives 1 and 2 were addressed using Confirmatory Factor Analysis (CFA) (Brown, 2015). CFA is a parametric technique with normality assumptions regarding data type and distribution (Brown, 2015; Byrne, 2010). The EPDS represents a challenge to these assumptions in a number of ways. Firstly, the four-point scoring of each item is more readily described as ordinal, ordered categorical or polytomous, in contrast to continuous or interval level data more commonly

associated with maximum-likelihood (ML) estimation techniques commonly used in CFA (Byrne, 2010; Kline, 2011). Secondly, though distributional characteristics of individual EPDS items are rarely published, scrutiny of individual EPDS item content indicates the potential for violation of the normal distribution criteria for parametric tests (C.R. Martin and Thompson, 2000). EPDS item 10. 'The thought of harming myself has occurred to me' has been identified as a 'suicide' question (Jomeen and Martin, 2005) and may therefore, given the general population of postpartum women, be comparatively less likely to be endorsed beyond a zero score. Importantly, advances in estimation methods in CFA have facilitated robust appraisal of models with data that is either non-normal or ordered categorical, or indeed both. An appropriate estimation method in these circumstances is the Weighted Least Squares with Means and Variances corrected (WLSMV) (Beauducel and Herzberg, 2006; Koziol and Bovaird; B. Muthén et al, 1997; B. O. Muthén, 1993). This estimation approach is also suitable for modest sample sizes (Brown, 2015; Flora and Curran, 2004; Jomeen and Martin, 2007).

The first model evaluated represented the empirical specification of the EPDS for clinical practice and as originally conceived by Cox et al. (1987), a uni-dimensional model of PND with all items loading on a single 'depression' factor. The second model tested was the uni-dimensional model of the EPDS comprising seven items (EPDS-7) proposed by (Gollan et al., 2017), with 'anxiety' items (Matthey, 2008) 3, 4 and 5 removed. The third model evaluated was a two-factor model of the EPDS comprising EPDS-7 items as a factor and the three 'anxiety' items as a second correlated factor (Gollan et al., 2017; Matthey, 2008; Phillips et al., 2009). The fourth model evaluated was the tri-dimensional model of Reichenheim et al. (2011) comprising three correlated factors of anhedonia (items 1, 2 and 6), anxiety (items 3, 4 and 5) and depression (items 7, 8, 9 and 10). This model was selected for evaluation due to the exhaustive statistical approach to model identification and evaluation conducted by Reichenheim and colleagues (2011). The fifth model evaluated was a modification of this model suggested by Reichenheim et al. (2011), a bi-factor model comprising a general factor and three uncorrelated specific factors (anhedonia, anxiety and depression). The

specific factors are specified by the same items as the three-factor correlated model (model 3). The sixth model (Tuohy and McVey, 2008) evaluated was a three-factor model comprising depression (items 7, 8, 9 and 10), anhedonia (items 3, 4 and 5) and anxiety symptoms (items 1 and 2). The seventh model was a three-factor model, based on a critique of the EPDS comprised a 7-item tri-dimensional structure (Matthey and Agostini, 2017), similar to that of Tuohy and McVey (2008) but with potentially ambiguous items removed (items 7 and 10). The eighth model evaluated was the three-factor 6-item model of Kozinszky et al (2017) comprising anhedonia (items 1 and 2), anxiety (items 4 and 5) and low mood (items 8 and 9). A unidimensional version of this model (model 9) was also evaluated to determine the relative contribution of multi-dimensionality over selection of what might be simply better performing items. Finally, the tri-dimensional model proposed by Coates et al. (2016) was evaluated comprising anhedonia (items 1 and 2), anxiety (items 3-6) and depression (items 7-10)

The ten models were evaluated by a variety of model fit indices (Bentler and Bonett, 1980). Specifically, the comparative fit index (CFI; (Bentler, 1990), the root mean squared error of approximation (RMSEA; (Steiger and Lind, 1980), and the weighted root mean residual (WRMR;(Yu, 2002) were used to evaluate model fit. CFI values > 0.90 indicates an acceptable fit (Hu and Bentler, 1995) more stringent CFI ≥ 0.95 indicating a good fit to the data (Hu and Bentler, 1999). RMSEA values ≤ 0.08 indicate acceptable model fit (Browne and Cudeck, 1993), and values of ≤0.05 indicative of good fit (Schumacker and Lomax, 2010). WRMR values of ≤0.10 are indicative of good model fit (Yu, 2002). The χ^2 statistic may be used to evaluate model fit, with a non-significant p value indicating acceptable model fit. However, the χ^2 is influenced by both sample size and data distribution, thus model evaluation is largely based on indices such as CFI and RMSEA rather than χ^2 (Byrne, 2010; Hooper et al, 2008; Vardavaki at al, 2015). Individual EPDS items were specified as ordered categorical within the analysis since not to do so can impact on model fit estimation (Gollan et al., 2017; Reichenheim et al., 2011). Objective 3 was addressed by comparing the best-fitting models of the EPDS identified from objectives 1 and 2 between the two observation points (three

months and six months). Measurement invariance evaluation requires the application of increasingly restrictive versions of the underlying model (Brown, 2015; C. R. Martin et al., 2017). Initially, datasets representing the two time points would be evaluated for model fit based on the previously established best-fit model. Following satisfactory model fit at each time point a configural invariance model would be estimated to determine consistency between time points of the factor model and pattern of loadings. Establishing configural invariance enables a more restrictive model to be evaluated where item-factor loadings are constrained to be equal across groups. Metric invariance is a pre-requisite to confirm equivalence of meaning of the underlying constructs specified within the model (Kline, 2011). Further constraints following observation of metric invariance can made, specifically the evaluation of a threshold model in which loadings and thresholds are constrained to be equal. Non-invariant items are identified by evidence of a significant difference between models (invariant/non-invariant), as evidenced by a reduction in CFI of >0.01 (Cheung and Rensvold, 2002) and an increase in RMSEA of >0.015 (Chen et al, 2008). Objective 4 was evaluated using the independent t-test to compare mean scores of the model EPDS total and sub-scale scores between the three month and six month datasets. It is predicted that there would be no statistically significant differences between mean scores. Objective 5 was evaluated by comparing the Cronbach (Cronbach, 1951) alpha of EPDS, EPDS-7 and sub-scales between the three month and six month data set using the statistical approach of Feldt et al (1987) and Diedenhofen and Musch (2016). It is predicted that no statistically significant differences would be observed between time points. Objective 6 was met by comparing the Pearson's r correlation coefficients of each combination of EPDS/EPDS sub-scales within a time point across both time points using the approach of Diedenhofen and Musch (2016) and Zou (2007). Objective 7 was evaluated by comparing case classification rates of both EPDS and EPDS-7 scores using the established thresholds for the EPDS (J.L. Cox and Holden (2003); J. L. Cox et al. (1987) and the thresholds for the EPDS-7 (Gollan et al., 2017).

Statistical analysis was conducted using the R programming language (R Core Team, 2017) and the R packages Lavaan (Rosseel, 2012), Cocron (Diedenhofen and Musch, 2016) and Cocor (Diedenhofen and Musch, 2015).

3. Results

3.1 Participant characteristics

A response rate of 28% was achieved with 504 women returning usable data to the pilot postal survey. Complete EPDS data was available on 484 participants (~4% missing data). Given that outliers can bias both estimation efficacy within a CFA (Yuan and Bentler, 2001) and deleteriously influence distributional normality (Brown, 2015), removal of outliers from the analysis is advised if the sample size permits (Meyers et al, 2006). The complete dataset was thus then screened for multivariate outliers by reference to Mahalanobis distances, and twenty-two outliers were found and excluded based on a distance from the centroid estimation of $\chi^2 > 29.59$. The final number of participants for which data was complete and multivariate normal was thus *N*=462 (~4.5% outliers from complete data). The mean age of participants was 32.02 (SD 5.60) years. The average duration of pregnancy was 39.06 (SD 2.39) weeks. The majority (*N*=446) of women (96%) had a single baby. The majority (*N*=417, 90%) of women had their baby in hospital. Two-hundred and thirty-six women (51%) had their baby delivered in either a midwifery-led unit or birth centre. Two hundred and twenty women (48%) were first-time mothers. Stratification of the complete prepared dataset (*N*=462) revealed *N*=245 women (53%) completed the EPDS at three months postpartum and *N*=217 women (47%) at six months postpartum. Younger women and those living in the most disadvantaged were less well represented as is common with national surveys of new mothers (Redshaw and Heikkila, 2010; Redshaw and Henderson, 2015). No statistically significant differences were observed between the two samples in terms of duration of pregnancy, age, those who had a single baby, those delivered in a midwifery-led unit or birth centre, those delivered in hospital or first-time

mother status. Salient details of the two samples and the results of the inferential comparison between groups are summarised in Table 1.

TABLE 1. ABOUT HERE

The mean item score and distributional characteristics of individual EPDS items is shown in Table 2. for each observation point. EPDS item 10 is shown to be highly skewed and kurtotic at both three months and six months postpartum.

TABLE 2. ABOUT HERE

3.2 Confirmatory factor analysis

The findings from each CFA evaluation at each time point and combined for each model are summarised in Table 3. The models of Tuohy and McVey (2008), the modified version of this model and the three-factor model of Kozinszky et al. (2017) were consistently observed to offer the best fit to data. Differences in fit indices between these three best-fit models were trivial. Multi-dimensional models were observed to offer better fit to data than single-factor models. The bi-factor model of Reichenheim et al. (2011) failed to yield a permissible factor solution and model fit in any of the datasets. Following a suggestion by one of the reviewers on an earlier version of this paper, we have rerun one of the models with outliers included to determine any impact on model fit. The rerun of the Tuohy and McVey (2008) model (all data) revealed a model fit similar to that found with outliers removed ($\chi^2_{(df=24)} = 35.79$, p =0.06, RMSEA = 0.03, WRMR = 0.53, CFI = 0.99, TLI = 0.99).

TABLE 3. ABOUT HERE

The best-fit models from the CFA were evaluated for measurement invariance and the findings summarised in Table 4. Using the ΔCFI criteria (Cheung and Rensvold, 2002) and ΔRMSEA (Chen et al., 2008) applied to increasingly restrictive models, no statistically significant measurement non-invariance was observed between the models tested across each time point.

TABLE 4. ABOUT HERE

3.3 EPDS mean score coherence

Comparisons between EPDS/EPDS sub-scale scores are summarised in Table 5. Statistically significant differences were observed between mean EPDS total, EPDS-7, the 'anxiety' sub-scale and the three-item anhedonia sub-scale with these scores being higher at six months. Effect sizes were however small.

TABLE 5. ABOUT HERE

3.4 Internal consistency evaluation

The findings from the internal consistency evaluation are summarised in Table 6. Cronbach alpha's for the EPDS/EPDS sub-scales were all acceptable based on threshold criteria of 0.70 (Nunnally, 1978) with the exception of the three-item anhedonia sub-scale. No statistically significant differences were observed between time points.

TABLE 6. ABOUT HERE

3.5 Equivalence of correlations between EPDS and EPDS sub-scales

The findings of the equivalence evaluation between correlations are summarised in Table 7. All correlations were statistically significant within each time point. No statistically significant differences were observed between correlations between EPDS/EPDS sub-scales between three month and six month observations.

TABLE 7. ABOUT HERE.

3.6 Case classification

The case classification rates for the EPDS and EPDS-7 are summarised in Table 8. It is noted descriptively that there are large discrepancies between classification rates using the EPDS conventional thresholds and the 'equivalent' thresholds (Gollan et al., 2017) for the EPDS-7. EPDS case classification using the 12/13 screening cut-off revealed *N*=65 (14%) case positive for the

complete dataset, *N*=26 (11%) case positive at three months and *N*=39 (18%) case positive at 6 months. A chi-square test of independence was conducted comparing the frequency of EPDS caseness (case positive) at each time point. A significant interaction was found ($\chi^2(1) = 5.16$, p = 0.02) reflecting the greater likelihood of EPDS caseness at six months. EPDS-7 case classification using the 4/5 screening cut-off suggested as equivalent to the 12/13 cut-off of the EPDS (Gollan et al., 2017) revealed *N*=140 (30%) case positive for the complete dataset, *N*=70 (29%) case positive at three months and *N*=70 (32%) case positive at 6 months. A chi-square test of independence was conducted comparing the frequency of EPDS-7 caseness at each time point and was revealed not to be statistically significant ($\chi^2(1) = 0.74$, $p = 0.39$).

TABLE 8. ABOUT HERE

4. Discussion

The findings from the current investigation address a number of aspects of the contemporary debate regarding not only the measurement characteristics of the EPDS, but also the use of the instrument in both clinical and non-clinical populations. The study is notable in using two independent datasets drawn using a common data capture methodology with three month and six month observation points. This approach has the advantage of being able to look at the replicability of the factor structure of the EPDS in a normal population sample, without the potential bias that may influence a repeated-measures design, for example idiosyncrasies of discrete aspects of a population deleteriously influencing findings from a psychometric analysis (Jomeen and Martin, 2007, 2008b).

A fundamental but important observation noted from the distributional characteristics of the EPDS items was the significant skew and kurtosis noted for item 10. ‘‘The thought of harming myself has occurred to me’. Noted for being a ‘suicide question’ (Jomeen and Martin, 2005), the issue of this item was also raised in a recent critique, due to potential ambiguity depending on when the questionnaire is administered (Matthey and Agostini, 2017). The skew and kurtosis of this item

noted in the current study represents an artefact of the low level of endorsement above zero, however, given that the datasets were screened for multivariate outliers which were consequently removed, it is important to consider what effect such an item may have both on total EPDS score in terms of total score deflation and also in terms of the appropriateness of applying of parametric approaches to data analysis assuming multivariate normal data. This is of particular relevance given that the vast majority of published studies on the EPDS do not report the distributional characteristics of the measure at an item level. It is also true that, although traditional parametric tests are established to be robust against violations of the assumptions that characterise their use (C.R. Martin and Thompson, 2000), the impact on structural equation models can be significant if data normality is assumed within the analysis but not realised within the data itself (Lubke and Muthén, 2004). Not accommodating the characteristics of the data into the model estimation method may directly lead to differing results in similarly specified models (Reichenheim et al., 2011).

Evaluation of model fit statistics revealed that the three-factor model of Tuohy and McVey (2008), the modified shorter version of this model and the three-factor model of Kozinszky et al. (2017) offered consistently the best fit to data in three month, six month and combined datasets. It was also clear from appraisal of all the models tested that the EPDS represents fundamentally a multi-dimensional measure, given the superiority of these conceptualisations of the measure over the uni-dimensional models evaluated. Further evaluation of the measurement invariance characteristics of the three best-fit models demonstrated clear invariance of the measurement models between the three month and six month datasets. This contrasts with the findings of Cunningham et al. (2015) who demonstrated a highly dynamic and changeable factor structure which as those authors themselves concede, could be strongly influenced by the treatment effects of intervention within a clinical group. Recognising the impact of treatment or population effects is critical to understanding the scope of use of a screening tool. These demarcation lines are not clear with EPDS studies, as evidenced by the contrasting results in factor stability observed between the current study and that of Cunningham et al. (2015). However, it is reassuring that a recent study of

postpartum African-American women (Lee King, 2012a) of low social economic status also found the best-fit model to be that of Tuohy and McVey (2008). Thus, given similarities with the best-fit model in the current study, the replicability of the underlying factor structure of the EPDS appears to be quite broad in general applicability, but not among more extreme clinical populations. A striking observation was the excellent fit of Kozinszky et al.'s (2017) three-factor model, not least because this 'theory-driven' model was determined by as being of particular relevance for antenatal screening (Kozinszky et al, 2017). However, it should be noted that though this model provided an excellent fit to our postnatal data, the unidimensional version of this model revealed a comparatively poorer fit, indeed, an extremely poor fit to data in terms of specific fit indices such as RMSEA and WRMR. This highlights the implicit contribution of multi-dimensionality to the excellent fit reported for the six-item three-factor model.

Surprisingly, the tri-dimensional model of Coates et al. (2016) did not offer the best fit to data. This was unexpected due to the consistency of good fit to data of this model found between antenatal and postnatal samples and also due to the large sample size utilised. There are at least three possibilities for the lack of exemplary fit of Coates et al.'s (2016) model to the current data: firstly, the model does offer a good fit to the data, but comparatively, it is not as good a fit; secondly, the estimation approach used (maximum-likelihood) might have not been the most suitable for the ordered categorical nature of EPDS data (Reichenheim et al., 2011) and thirdly, the translation of a model derived from exploratory factor analysis (EFA) might not be the most convincing model when applied to CFA. This highlights a potential limitation of EFA approaches which often use 'rule of thumb' to determine both the number of appropriate factors and the threshold for item-factor loadings. Coates et al. (2016) themselves highlight potential issues of cross-loading with item 6 and (Kozinszky et al., 2017) emphasise the lack of acceptability of a uni-dimensional model of the EPDS that was observed in their initial EFA.

The application of the recent promising development of the EPDS-7 (Gollan et al., 2017) yielded some unexpected findings in the CFA, in particular, the superior fit of a two-factor 10-item model comprising the EPDS-items correlated with a factor comprising the three 'anxiety' items, in contrast to Gollan et al. (2017) who found the reverse and also utilised their findings in this regard, in part, for the justification of the EPDS-7. Further scrutiny of the EPDS-7 factor model reveals that in both the combined dataset and the three month dataset, the model fails to reach acceptability according to the RMSEA fit measure, as was also reported in the EPDS-7 validation paper (Gollan et al, 2017).

Corroboration of the relative measurement intransigence of the best-fit models of the EPDS observed in the CFA's and invariance evaluations can be found in the findings of no statistically significant differences observed between internal consistency estimations between the two time points and between scale/sub-scale correlations between the two time points. Taken together, the findings thus far confirm a tri-dimensional structure that represents the data extremely well and is consistent across time points. These findings offer a robust context for the comparison of mean scale/sub-scale scores between three months and six months. It was observed that for EPDS total, EPDS-7 and three-item anhedonia sub-scale scores that mean scores were significantly higher at the six month time point. The lack of variation in EPDS structure highlighted above supports the view that these represent real differences rather than 'method effects' and thus demonstrates the salient but often neglected area of considering and agreeing when is the most appropriate time to screen for PND. Given that the postnatal period is defined to last 12 months and screening practice varies in terms of timing varies hugely, these findings suggest that not only is a consistent screening window or timeframe vital, but also the suggestion of retesting should be seriously considered. Matthey and Agostini (2017) have highlighted convincing evidence for transient experience of depressive symptomology postpartum, however, the findings from the current study including the observation of significantly proportions of women screening positive between three and six month observation

points would also suggest that even a relatively late screen (three months) would not detect a proportion of those who would screen positive at six months.

A significant issue that has been raised regarding the EPDS is the impact of culture, socio-economic diversity and availability of resources to screen on the selection of threshold cut-off scores (Lee King, 2012a, 2012b; Matthey and Agostini, 2017). It was observed in the current investigation that 'equivalent' threshold cut-off scores for the EPDS and EPDS-7 resulted in wide variation in case identification rates and statistically significant different classification profiles for each measure. This observation would suggest that the EPDS and EPDS-7 are not equivalent in case classification utility and thus the potential for false negative case classification is increased. While it is acknowledged that this finding could be influenced by differences in the populations under investigation, it is should also be noted that Gollan et al. (2017) highlighted the equivalence of cut-offs between versions, and therefore it would be presumed that these would translate measurement-wise between populations. Thus a population with for example atypical low or high scores would still screen at the same classification rates on either version of the tool. This was not found to be the case in the present study. A further complication regarding scoring cut-off thresholds is the perennial and currently unaddressed issue regarding the EPDS is the tension between the evidence of clear multidimensionality within the measure, and the application of a cut-score which assumes unidimensionality. If the potential benefit of maximising on screening efficacy of the EPDS through the use of screening cut-off scores while accepting the measure is multidimensional, then it would be address the tension highlighted if threshold-scores could be reliably determined for EPDS sub-scales, notably within the three 'best-fit' models observed within the current investigation. This would of course require a further study and an evaluation of these embedded EPDS sub-scale scores against a 'gold standard' clinical diagnosis and evaluation of the receiver-operating characteristic against diagnosis. Firstly, the screening accuracy of PND of EPDS sub-scales could be evaluated against the full EPDS itself, since obviously the sub-scales are embedded within the measure. Secondly, under the rubric of using a full differential diagnosis as a gold standard and assessing for

other significant postpartum mental health concerns such as generalized anxiety disorder, embedded EPDS sub-scales which feature more specifically anxiety-related content may be evaluated to determine additional screening value and utility.

The study had a number of limitations. Firstly, our response rate was modest at 28%. It is therefore entirely possible that there may be inherent bias in those that returned the questionnaires which may impact of the representativeness of the sample compared to the population. A second limitation is that though we looked at the psychometric performance of the EPDS in two samples, we did not conduct a longitudinal study to determine the change over time in the same participant group. A logical progression from the current study would be a replication study using a longitudinal design to determine within-group variability to complement the between-group variability investigated here. Given the consistency of the between-groups observations in the current study, it would be surprising to find within-groups variability in a follow-up study being more that that currently observed, indeed, it would be anticipated to be less, however, the plausibility of such an assumption remains to be evaluated and would provide useful additional evidence that would contribute to the debate over the measurement characteristics of the EPDS.

In conclusion, the current investigation has found the best-fit to data for the EPDS to be three-factor models, specifically, Tuohy and McVey (2008), a modified version of this model and Kozinszky et al.'s (2017) three-factor model. We found these models not only to be a good fit and replicable between two postpartum datasets at three month and six month observation points, but also the factor structures to be invariant, thus engendering confidence in the measurement veracity of sub-scale scores that may be derived from this model. Finally, the study also raised concerns regarding the measurement and case identification characteristics of the EPDS derivative, the EPDS-7.

Acknowledgements

We would particularly like to thank the women who took part in this research by completing the questionnaire. We also wish to thank staff at the Office for National Statistics who drew the sample and managed the mailings and Sian Harrison who managed the survey more broadly, Ciconi who printed and prepared the survey packs and were responsible for the data entry and Qualtrics who set up the online survey.

This paper reports on an independent study which is funded by the Policy Research Programme in the Department of Health. The views expressed in this report are those of the authors and do not necessarily reflect the views of the Department.

Conflict of interest

The authors declare they have no conflict of interest.

References

American College of Obstetricians and Gynecologists' Committee on Obstetric Practice, 2015. The American College of Obstetricians and Gynecologists Committee Opinion no. 630. Screening for perinatal depression. Obstet. Gynecol., 125, 1268-1271.

Banti, S., Mauri, M., Oppo, A., Borri, C., Rambelli, C., Ramacciotti, G. et al, 2011. From the third month of pregnancy to 1 year postpartum. Prevalence, incidence, recurrence, and new onset of depression. Results from the perinatal depression-research & screening unit study. Compr. Psychiatry 52(4), 343-351. doi: 10.1016/j.comppsych.2010.08.003

Beauducel, A., Herzberg, P. Y., 2006. On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. Struct. Equat. Modeling: Multidisc. J. 13(2), 186-203. doi: 10.1207/s15328007sem1302_2

Bentler, P. M., 1990. Comparative fit indexes in structural models. Psychol. Bull. 107(2), 238-246.

Bentler, P. M., Bonett, D. G., 1980. Significance tests and goodness of fit in the evaluation of covariance structures. Psychol. Bull. 88, 588-606.

Brouwers, E. P., van Baar, A. L., Pop, V. J., 2001. Does the Edinburgh Postnatal Depression Scale measure anxiety? J. Psychosom. Res. 51(5), 659-663.

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.

Cameron, E. E., Hunter, D., Sedov, I. D., Tomfohr-Madsen, L. M., 2017. What do dads want? Treatment preferences for paternal postpartum depression. J. Affect. Disord. 215, 62-70. doi: 10.1016/j.jad.2017.03.031

Chabrol, H., Teissedre, F., 2004. Relation between Edinburgh Postnatal Depression Scale scores at 2–3 days and 4–6 weeks post-partum. J. Reprod. Infant. Psychol. 22, 33-39.

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. Sociol. Methods Res. 36, 462–494.

Cheung, G. W., Rensvold, R. B., 2002. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct. Equat. Modeling: Multidisc. J. 9(2), 233-255. doi: 10.1207/S15328007SEM0902_5

Clark, L. A., Watson, D., 1991. Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. J. Abnorm. Psychol. 100(3), 316-336.

Coates, R., Ayers, S., de Visser, R., 2016. Factor structure of the Edinburgh Postnatal Depression Scale in a population-based sample. Psychological Assessment. doi: 10.1037/pas0000397

Cox, J. L., Holden, J. M., 2003. A Guide to the Edinburgh Postnatal Depression Scale. London: Gaskell.

Cox, J. L., Holden, J. M., Sagovsky, R., 1987. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. Br. J. Psychiatry 150, 782-786.

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

Cunningham, N. K., Brown, P. M., Page, A. C., 2015. Does the Edinburgh Postnatal Depression Scale measure the same constructs across time? Arch. Womens Ment. Health 18(6), 793-804. doi: 10.1007/s00737-014-0485-9

Dahlen, H. G., Barnett, B., Kohlhoff, J., Drum, M. E., Munoz, A. M., Thornton, C., 2015. Obstetric and psychosocial risk factors for Australian-born and non-Australian born women and associated pregnancy and birth outcomes: a population based cohort study. BMC Pregnancy Childbirth 15, 292. doi: 10.1186/s12884-015-0681-2

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

Diedenhofen, B., Musch, J., 2016. cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. Int. J. Internet Sci. 11(1), 51–60.

Fairbrother, N., Woody, S. R., 2008. New mothers' thoughts of harm related to the newborn. Arch. Womens Ment.Health 11(3), 221-229. doi: 10.1007/s00737-008-0016-7

Feldt, L. S., Woodruff, D. J., Salih, F. A., 1987. Statistical inference for coefficient alpha. Applied Psychol. Measurement 11, 93-103.

Flora, D. B., Curran, P. J., 2004. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychol. Methods 9(4), 466-491. doi: 10.1037/1082-989X.9.4.466

Gollan, J. K., Wisniewski, S. R., Luther, J. F., Eng, H. F., Dills, J. L., Sit, D., et al, 2017. Generating an efficient version of the Edinburgh Postnatal Depression Scale in an urban obstetrical population. J. Affect. Disord. 208, 615-620. doi: 10.1016/j.jad.2016.10.013

Hartley, C. M., Barroso, N., Rey, Y., Pettit, J. W., Bagner, D. M., 2014. Factor structure and psychometric properties of English and Spanish versions of the Edinburgh Postnatal Depression Scale among Hispanic women in a primary care setting. J. Clin. Psychol. 70(12), 1240-1250. doi: 10.1002/jclp.22101

Hooper, D., Coughlan, J., Mullen, M. R., 2008. Structural equation modelling: guidelines for determining model fit. Electronic J. Business Res. Meth. 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. Struct. Equat. Modeling 6, 1-55.

Jennings, K. D., Ross, S., Popper, S., Elmore, M., 1999. Thoughts of harming infants in depressed and nondepressed mothers J. Affect. Disord. 54(1-2), 21-28.

Jomeen, J., Martin, C. R., 2005. Confirmation of an occluded anxiety component within the Edinburgh Postnatal Depression Scale (EPDS) during early pregnancy. J. Reprod. Infant Psychol. 23(2), 143–154.

Jomeen, J., Martin, C. R., 2007. Replicability and stability of the multidimensional model of the Edinburgh Postnatal Depression Scale in late pregnancy. J. Psychiat. Ment. Health Nurs. 14(3), 319-324. doi: 10.1111/j.1365-2850.2007.01084.x

Jomeen, J., Martin, C. R., 2008a. The impact of choice of maternity care on psychological health outcomes for women during pregnancy and the postnatal period. J. Eval. Clin. Pract. 14(3), 391-398. doi: 10.1111/j.1365-2753.2007.00878.x

Jomeen, J., Martin, C. R., 2008b. Reflections on the notion of post-natal depression following examination of the scoring pattern of women on the EPDS during pregnancy and in the post-natal period. J. Psychiatry. Ment. Health Nurs. 15(8), 645-648. doi: 10.1111/j.1365-2850.2008.01282.x

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

Kozinszky, Z., Toreki, A., Hompoth, E. A., Dudas, R. B., Nemeth, G., 2017. A more rational, theory-driven approach to analysing the factor structure of the Edinburgh Postnatal Depression Scale. Psychiatry Res. 250, 234-243. doi: 10.1016/j.psychres.2017.01.059

Koziol, N. A., Bovaird, J. A., 2017. The impact of model parameterization and estimation methods on tests of measurement invariance with ordered polytomous data. Educ. Psychol. Measurement 78(2), 272-296. 0013164416683754. doi: doi:10.1177/0013164416683754

Lee King, P. A., 2012a. Replicability of structural models of the Edinburgh Postnatal Depression Scale (EPDS) in a community sample of postpartum African American women with low socioeconomic status. Arch. Womens Ment. Health 15(2), 77-86. doi: 10.1007/s00737-012-0260-8

Lee King, P. A., 2012b. Validity of postpartum depression screening across socioeconomic groups: a review of the construct and common screening tools. J. Health Care Poor Underserved 23(4), 1431-1456. doi: 10.1353/hpu.2012.0163

Lubke, G. H., Muthén, B. O., 2004. Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. Struct. Equat. Modeling 11(4), 514-534.

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 Birth 30(4):e172-e178. doi: 10.1016/j.wombi.2016.11.006

Martin, C. R., Thompson, D. R., 2000. Design and Analysis of Clinical Nursing Research Studies. London: Routledge.

Matthey, S. 2008. Using the Edinburgh Postnatal Depression Scale to screen for anxiety disorders. Depress. Anxiety 25(11), 926-931. doi: 10.1002/da.20415

Matthey, S., Agostini, F., 2017. Using the Edinburgh Postnatal Depression Scale for women and men-some cautionary thoughts. Arch. Womens Ment. Health 20(2), 345-354. doi: 10.1007/s00737-016-0710-9

Meades, R., Ayers, S., 2011. Anxiety measures validated in perinatal populations: a systematic review. J. Affect. Disord. 133(1-2), 1-15. doi: 10.1016/j.jad.2010.10.009

Meyers, L. S., Gamst, G., Guarino, A. J., 2006. Applied Multivariate Research Design and Interpretation. Thousand Oaks, CA Sage.

Milgrom, J., Holt, C., 2014. Early intervention to protect the mother-infant relationship following postnatal depression: study protocol for a randomised controlled trial. Trials 15, 385. doi: 10.1186/1745-6215-15-385

Moraes, G. P., Lorenzo, L., Pontes, G. A., Montenegro, M. C., Cantilino, A., 2017. Screening and diagnosing postpartum depression: when and how? Trends Psychiatry Psychother. 39(1), 54-61. doi: 10.1590/2237-6089-2016-0034

Morgan, G. B., Hodge, K. J., Wells, K. E., Watkins, M. W., 2015. Are fit indices biased in favor of bi-factor models in cognitive ability research? : a comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. J. Intelligence 3, 2-20. doi: 10.3390/jintelligence301000

Muthén, B., du Toit, S. H. C., Spisic, D., 1997. Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Los Angeles, CA. https://www.statmodel.com/download/Article_075.pdf

Muthén, B. O., 1993. Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), Testing Structural Equation Models (pp. 205–243). Newbury Park, CA: Sage.

National Institute for Health and Care Excellence, 2015. Antenatal and postnatal mental health: clinical management and service guidance (cg192). http://www.nice.org.uk/guidance/cg192.

Nunnally, J., 1978. Psychometric Theory. New York: McGraw-Hill.

O'Hara, M. W., Swain, A. M., 1996. Rates and risk of postpartum depression—a meta-analysis. International Rev. Psychiatry 8(1), 37-54. doi: 10.3109/09540269609037816

Pallant, J. F., Miller, R. L., Tennant, A., 2006. Evaluation of the Edinburgh post natal depression scale using Rasch analysis. BMC Psychiatry 6, 28. doi: 10.1186/1471-244X-6-28

Phillips, J., Charles, M., Sharpe, L., Matthey, S., 2009. Validation of the subscales of the Edinburgh Postnatal Depression Scale in a sample of women with unsettled infants. J. Affect. Disord. 118(1-3), 101-112. doi: 10.1016/j.jad.2009.02.004

Pope, C. J., Xie, B., Sharma, V., Campbell, M. K., 2013. A prospective study of thoughts of self-harm and suicidal ideation during the postpartum period in women with mood disorders. Arch. Women's Ment. Health 16(6), 483-488. doi: 10.1007/s00737-013-0370-y

Redshaw, M., Heikkila, K., 2010. Delivered with care: a national survey of women's experience of maternity care. NPEU, Oxford.

Redshaw, M., Henderson, J., 2015. Safely delivered: a national survey of women's experience of maternity care. NPEU, Oxford.

Reichenheim, M. E., Moraes, C. L., Oliveira, A. S., Lobato, G., 2011. Revisiting the dimensional structure of the Edinburgh Postnatal Depression Scale (EPDS): empirical evidence for a general factor. BMC Med. Res. Methodol. 11, 93. doi: 10.1186/1471-2288-11-93

Ross, L. E., Gilbert Evans, S. E., Sellers, E. M., Romach, M. K., 2003. Measurement issues in postpartum depression part 1: anxiety as a feature of postpartum depression. Arch. Women's Ment. Health 6(1), 51-57. doi: 10.1007/s00737-002-0155-1

Rosseel, Y., 2012. Lavaan: An R package for structural equation modeling. J. Statist. Software 48(2), 1-36.

Schumacker, R. E., Lomax, R. G., 2010. A beginner's guide to structural equation modelling (3rd ed.). New York: Routledge/Taylor and Francis Group.

Smith, E. K., Gopalan, P., Glance, J. B., Azzam, P. N., 2016. Postpartum depression screening: a review for psychiatrists. Harv. Rev. Psychiatry 24(3), 173-187. doi: 10.1097/HRP.0000000000000103

Steiger, J. H., Lind, J., 1980. Statistically-based tests for the number of common factors. Annual Spring Meeting of the Psychometric Society, Iowa City, USA.

Team, R. C., 2017. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Teissedre, F., Chabrol, H., 2004. Detecting women at risk for postnatal depression using the Edinburgh Postnatal Depression Scale at 2 to 3 days postpartum. Can. J. Psychiatry. Revue Can. Psychiatrie 49(1), 51-54. doi: 10.1177/070674370404900108

Tohotoa, J., Maycock, B., Hauck, Y. L., Dhaliwal, S., Howat, P., Burns, S., et al, 2012. Can father inclusive practice reduce paternal postnatal anxiety? A repeated measures cohort study using the Hospital Anxiety and Depression Scale. BMC Pregnancy Childbirth 12, 75. doi: 10.1186/1471-2393-12-75

Tuohy, A., McVey, C., 2008. Subscales measuring symptoms of non-specific depression, anhedonia, and anxiety in the Edinburgh Postnatal Depression Scale. Br. J. Clin. Psychol. 47(Pt 2), 153-169. doi: 10.1348/014466507X238608

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). J. Reprod. Infant Psychol. 33(5), 488-503. doi: 10.1080/02646838.2015.1035235

Wisner, K. L., Sit, D. K., McShea, M. C., Rizzo, D. M., Zoretich, R. A., Hughes, C. L., et al, 2013. Onset timing, thoughts of self-harm, and diagnoses in postpartum women with screen-positive depression findings. JAMA Psychiatry 70(5), 490-498. doi: 10.1001/jamapsychiatry.2013.87

Yu, C. Y., 2002. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. (Doctoral dissertation), University of California, Los Angeles.

Yuan, K.-H., Bentler, P. M., 2001. Effect of outliers on estimators and tests in covariance structure analysis. Br. J. Math. Statistical Psychol. 54 ,161–175. doi:10.1348/000711001159366

Zigmond, A. S., Snaith, R. P., 1983. The hospital anxiety and depression scale. Acta Psychiatr. Scand. 67(6), 361-370.

Zou, G. Y., 2007. Toward using confidence intervals to compare correlations. Psychol. Methods 12, 399–413. doi: doi: 10.1037/1082-989X.12.4.399

Table 1.

Comparison of background variables between the three month and six month postpartum samples and results of inferential statistical testing.

Variable	Sample		Analysis	
	Three month postpartum (*N*=245)	Six month postpartum (*N*=217)	Test statistic	*p*
Age (years)	31.70 (5.93)	32.39 (5.20)	$t_{(454)} = 1.31$	0.19
Pregnancy Duration (weeks)	38.87 (2.44)	39.29 (2.31)	$t_{(448)} = 1.88$	0.06
Single baby (yes/no)	234/10	212/5	$\chi^2_{(df=1)} = 0.67$	0.41
Midwifery-led unit/birth centre	118/123	118/97	$\chi^2_{(df=1)} = 1.37$	0.24
First time mother (yes/no)	111/134	109/107	$\chi^2_{(df=1)} = 1.03$	0.31
Hospital born (yes/no)	223/18	194/21	$\chi^2_{(df=1)} = 0.50$	0.48

Table 2.

Mean, standard deviation and distributional characteristics of EPDS items as a function of time point of observation (months postpartum). se = standard error of kurtosis.

Item	Time point	Mean	SD	Skew	Kurtosis	se
EPDS1.	Three	0.20	0.43	1.98	3.06	0.03
EPDS2.	Three	0.23	0.50	2.27	5.44	0.03
EPDS3.	Three	1.40	0.88	0.08	-0.72	0.06
EPDS4.	Three	1.16	0.97	0.24	-1.08	0.06
EPDS5.	Three	0.75	0.89	0.79	-0.62	0.06
EPDS6.	Three	1.24	0.90	-0.01	-1.04	0.06
EPDS7.	Three	0.29	0.63	2.23	4.51	0.04
EPDS8.	Three	0.63	0.79	0.86	-0.57	0.05
EPDS9.	Three	0.44	0.66	1.62	2.91	0.04
EPDS10.	Three	0.04	0.20	4.61	19.36	0.01
EPDS1.	Six	0.28	0.52	1.67	1.91	0.03
EPDS2.	Six	0.29	0.54	1.68	1.89	0.04
EPDS3.	Six	1.50	0.90	-0.12	-0.80	0.06
EPDS4.	Six	1.33	0.95	-0.11	-1.11	0.06
EPDS5.	Six	0.87	0.96	0.71	-0.68	0.06
EPDS6.	Six	1.45	0.82	-0.33	-0.64	0.06
EPDS7.	Six	0.53	0.80	1.38	0.96	0.05
EPDS8.	Six	0.80	0.86	0.52	-1.10	0.06
EPDS9.	Six	0.43	0.63	1.16	0.23	0.04
EPDS10.	Six	0.04	0.24	6.33	42.38	0.02

Table 3.

Confirmatory factor analysis of measurement models for 3 month, 6 month and combined data.

Model	N factors	N items	Time	WLSMV χ^2	df	*p*	RMSEA (95% CI)	WRMR	CFI	TLI
Cox et al. (1987)	1	10	All data	274.99	35	<0.001	0.12 (0.11-0.14)	1.63	0.93	0.91
Gollan et al. (2017)	1	7	All data	79.76	14	<0.001	0.10 (0.08-0.12)	1.14	0.97	0.96
Gollan et al. (2017)	2	10	All data	124.93	34	<0.001	0.08 (0.06-0.09)	1.05	0.97	0.97
Reichenheim et al. (2011)	3	10	All data	79.82	32	<0.001	0.06 (0.04-0.07)	0.80	0.99	0.98
Tuohy and McVey (2008)	3	9	All data	34.10	24	0.08	0.03 (0.01-0.05)	0.54	0.99	0.99
Modified three-factor	3	7	All data	18.30	11	0.08	0.04 (0.01-0.07)	0.46	0.99	0.99
Kozinszky et al. (2017)	3	6	All data	6.71	6	0.35	0.02 (0.01-0.06)	0.28	0.99	0.99
Kozinszky et al. (2017)	1	6	All data	168.23	9	<0.001	0.20 (0.17-0.22)	2.10	0.92	0.87
Coates et al. (2016)	3	10	All data	138.71	32	<0.001	0.09 (0.07-0.10)	1.06	0.97	0.96
Cox et al. (1987)	1	10	3 month	145.07	35	<0.001	0.11 (0.10-0.13)	1.16	0.94	0.92
Gollan et al. (2017)	1	7	3 month	55.76	14	<0.001	0.11 (0.08-0.14)	0.98	0.96	0.94
Gollan et al. (2017)	2	10	3 month	78.68	34	<0.001	0.07 (0.05-0.09)	0.82	0.97	0.97
Reichenheim et al. (2011)	3	10	3 month	48.61	32	0.03	0.05 (0.02-0.07)	0.60	0.99	0.99
Tuohy and McVey (2008)	3	9	3 month	27.53	24	0.28	0.03 (0.01-0.06)	0.46	0.99	0.99
Modified three-factor	3	7	3 month	12.57	11	0.32	0.02 (0.01-0.07)	0.37	0.99	0.99
Kozinszky et al. (2017)	3	6	3 month	6.45	6	0.37	0.02 (0.01-0.09)	0.27	0.99	0.99
Kozinszky et al. (2017)	1	6	3 month	73.00	9	<0.001	0.17 (0.14-0.21)	1.37	0.94	0.90
Coates et al. (2016)	3	10	3 month	74.78	32	<0.001	0.07 (0.05-0.10)	0.75	0.98	0.97
Cox et al. (1987)	1	10	6 month	158.36	35	<0.001	0.13 (0.11-0.15)	1.29	0.93	0.91
Gollan et al. (2017)	1	7	6 month	34.95	14	0.001	0.08 (0.05-0.12)	0.74	0.98	0.97
Gollan et al. (2017)	2	10	6 month	74.64	34	<0.001	0.07 (0.05-0.10)	0.83	0.98	0.97
Reichenheim et al. (2011)	3	10	6 month	61.83	32	0.001	0.07 (0.04-0.09)	0.73	0.98	0.98
Tuohy and McVey (2008)	3	9	6 month	30.94	24	0.16	0.04 (0.01-0.07)	0.53	0.99	0.99
Modified three-factor	3	7	6 month	19.57	11	0.05	0.06 (0.01-0.10)	0.49	0.99	0.99
Kozinszky et al. (2017)	3	6	6 month	8.25	6	0.21	0.04 (0.01-0.10)	0.31	0.99	0.99
Kozinszky et al. (2017)	1	6	6 month	107.03	9	<0.001	0.23 (0.19-0.26)	1.71	0.91	0.85
Coates et al. (2016)	3	10	6 month	91.92	32	<0.001	0.09 (0.07-0.12)	0.92	0.97	0.95

Table 4.

Measurement invariance evaluation of best-fit three-factor correlated models.

Model	Scaled χ^2 (df)	$\Delta\chi^2$	Δdf	Δ*p*	CFI	ΔCFI	RMSEA	ΔRMSEA
Tuohy and McVey (2008)								
Configural	28.82 (48)				0.996		0.031	
Loadings	43.21 (54)	16.34	6	0.01	0.991	0.005	0.045	0.014
Threshold	47.61 (64)	5.43	10	0.86	0.993	0.002	0.036	0.009
Means	65.42 (67)	6.96	3	0.07	0.989	0.005	0.046	0.010
Modified three-factor								
Configural	14.58 (22)				0.996		0.046	
Loadings	18.28 (26)	6.94	4	0.14	0.995	0.001	0.046	0.000
Threshold	22.44 (34)	7.12	8	0.52	0.996	0.001	0.036	0.010
Means	36.44 (37)	5.54	3	0.14	0.992	0.004	0.049	0.013
Kozinszky et al. (2017)								
Configural	5.17 (12)				0.999		0.031	
Loadings	9.61 (15)	10.59	3	0.01	0.995	0.003	0.053	0.022
Threshold	12.36 (21)	5.69	6	0.46	0.997	0.001	0.039	0.014
Means	23.32 (24)	5.39	3	0.15	0.993	0.004	0.052	0.013

Table 5.

Comparison of EPDS total score, EPDS-7 total score and Reichenheim et al. (2011), Tuohy and McVey (2008) model and modified model sub-scale scores, and Coates et al. (2016) anxiety sub-scale at three month and six month observation points. Standard deviations are in parentheses, degrees of freedom = 460, CI = confidence interval, ES = effect size.

EPDS scale/	Three months postpartum (*N*=245)	Six months postpartum (*N*=217)	95% CI	*t*	*p*	Hedges g	H' g 95% CI	ES sub-scale
Anhedonia (3-item)[+]	1.67 (1.47)	2.01 (1.49)	-0.62 to -0.07	2.50	0.01	-0.23	-0.42 to -0.05	Small
Anhedonia (2-item)[*]	0.43 (0.83)	0.57 (0.94)	-0.30 to 0.02	1.67	0.09	-0.16	-0.34 to 0.03	Neglig.
Anxiety (3-item)[+]	3.31 (2.26)	3.70 (2.27)	-0.81 to 0.03	1.83	0.07	-0.17	-0.35 to 0.01	Neglig.
Anxiety (2-item)[^]	1.91 (1.69)	2.20 (1.72)	-0.60 to 0.02	1.81	0.07	-0.17	-0.35 to 0.01	Neglig.
Anxiety (4-item)[†]	4.55 (2.85)	5.14 (2.77)	-1.11 to -0.08	2.27	0.02	-0.21	-0.40 to -0.03	Small
Depression (4-item)[#]	1.40 (1.84)	1.80 (2.02)	-0.75 to -0.04	2.19	0.03	-0.20	-0.39 to -0.02	Small
Depression (2-item)[*]	1.07 (1.31)	1.23 (1.34)	-0.40 to 0.08	1.30	0.19	-0.12	-0.30 to 0.06	Neglig.
EPDS-7	3.07 (2.96)	3.81 (3.20)	-1.30 to -0.17	2.57	0.01	-0.24	-0.42 to -0.06	Small
EPDS Total	6.38 (4.74)	7.51 (4.87)	-2.01 to -0.25	2.52	0.01	-0.23	-0.42 to -0.05	Small

Note: To control for type 1. error *p* criteria for statistical significance set at a more conservative 0.01.

Sub-scales: [+]Reichenheim et al. (2011); [#]Tuohy and McVey (2008); [*]Modified model based on Tuohy and McVey (2008); [†]Coates et al. (2016); [^]Kozinszky et al. (2017); EPDS-7 (Gollan et al., 2017); Total score (Cox et al., 1987).

Items: Anhedonia three-item (1,2,6); : Anhedonia two-item (1,2); Anxiety three-item (3,4,5); Anxiety two-item (4,5); Anxiety four-item (3,4,5,6); Depression four-item (7,8,9,10); Depression two-item (8,9); EPDS-7 (1,2,6,7,8,9,10).

Table 6.

Comparison of EPDS total score, EPDS-7 and Reichenheim et al. (2011), Tuohy and McVey's (2008) model and modified model Coates et al. (2016) model sub-scale Cronbach's alpha at three months and six months postpartum (df = 1).

Scale	Three month	Six month	χ^2	*p*
EPDS total	0.85	0.84	0.20	0.66
Anhedonia (3-item)	0.64	0.67	0.22	0.64
Anhedonia (2-item)	0.75	0.74	0.03	0.86
Anxiety (3-item)	0.78	0.74	0.80	0.37
Anxiety (2-item)	0.78	0.78	0.01	0.98
Depression (4-item)*	0.74	0.74	0.01	0.98
Depression (2-item)	0.77	0.75	0.13	0.72
EPDS-7	0.80	0.82	0.48	0.49
Anxiety (4-item)	0.79	0.76	0.61	0.43

Note: Calculated to three decimal points for statistical comparison purposes.

Table 7.

Correlations of EPDS total score, EPDS-7 and Reichenheim et al. (2011), Tuohy and McVey's (2008) model and modified model and Coates et al. (2016) sub-scales at three months and six months postpartum.

Scale combination	Three month *r*	Six month *r*	Z	95% CI	*p*
EPDS - Anhedonia (3-item)	0.81	0.80	0.30	(-0.05 to 0.08)	0.76
EPDS - Anhedonia (2-item)	0.645	0.65	0.17	(-0.12 to 0.10)	0.87
EPDS – Anxiety (3-item)	0.87	0.84	1.19	(-0.02 to 0.08)	0.23
EPDS - Depression (4-item)	0.85	0.87	0.82	(-0.07 to 0.03)	0.41
EPDS - Depression (2-item)	0.830	0.83	0.07	(-0.06 to 0.06)	0.95
EPDS - EPDS-7	0.93	0.92	0.74	(-0.06 to 0.04)	0.46
Anhedonia-Anxiety (3-item)	0.55	0.47	1.15	(-0.06 to 0.22)	0.25
Anhedonia-Depression (4-item)	0.60	0.65	0.88	(-0.16 to 0.06)	0.38
Anxiety (2-item) -Depression (4-item)	0.56	0.55	0.15	(-0.12 to 0.14)	0.88
Anhedonia (2-item)-Anhedonia (3-item)	0.83	0.87	1.54	(-0.09 to 0.01)	0.12
Anhedonia (2-item)-Anxiety (3-item)	0.41	0.32	1.11	(-0.07 to 0.25)	0.27
Anhedonia (2-item)-Depression (4-item)	0.49	0.57	1.19	(-0.21 to 0.05)	0.23
Depression (2-item)-Anhedonia (3-item)	0.60	0.63	0.51	(-0.14 to 0.09)	0.61
Depression (2-item)-Anhedonia (2-item)	0.48	0.55	1.02	(-0.20 to 0.07)	0.31
Depression (2-item)-Anxiety (3-item)	0.57	0.52	0.76	(-0.08 to 0.18)	0.45
Depression (2-item)-Depression (4-item)	0.95	0.94	1.00	(-0.01 to 0.03)	0.32
EPDS-7 - Anhedonia (3-item)	0.87	0.88	0.45	(-0.05 to 0.03)	0.65
EPDS-7 - Anhedonia (2-item)	0.71	0.77	1.42	(-0.14 to 0.02)	0.16
EPDS-7 – Anxiety (3-item)	0.62	0.57	0.83	(-0.07 to 0.17)	0.41
EPDS-7 - depression (4-item)	0.92	0.94	1.59	(-0.05 to 0.01)	0.11
EPDS-7 - depression (2-item)	0.89	0.89	0.15	(-0.04 to 0.04)	0.88
Anxiety (2-item) - EPDS	0.84	0.80	1.31	(-0.02 to 0.10)	0.19
Anxiety (2-item) - EPDS-7	0.61	0.56	0.81	(-0.07 to 0.17)	0.42
Anxiety (2-item) - Anhedonia (3-item)	0.55	0.44	1.56	(-0.03 to 0.25)	0.12
Anxiety (2-item) - Anhedonia (2-item)	0.43	0.32	1.37	(-0.05 to 0.27)	0.17
Anxiety (2-item) - Anxiety (3-item)	0.95	0.94	1.00	(-0.01 to 0.03)	0.32
Anxiety (2-item) - Depression (4-item)	0.55	0.55	0.03	(-0.13 to 0.13)	0.98
Anxiety (2-item) - Depression (2-item)	0.55	0.51	0.59	(-0.09 to 0.17)	0.55

Table 7 (continued). Correlations of EPDS total score, EPDS-7 and Reichenheim et al. (2011), Tuohy and McVey (2008) and modified model sub-scales at three months and six months postpartum.

Scale combination	Three month *r*	Six month *r*	Z	95% CI	*p*
Anxiety (4-item) - EPDS	0.93	0.90	1.98	(0.01 to 0.06)	0.05
Anxiety (4-item) - EPDS-7	0.74	0.68	1.29	(-0.03 to 0.15)	0.20
Anxiety (4-item) - Anhedonia (3-item)	0.71	0.63	1.55	(-0.02 to 0.18)	0.12
Anxiety (4-item) - Anhedonia (2-item)	0.47	0.39	1.05	(-0.07 to 0.23)	0.30
Anxiety (4-item) - Anxiety (3-item)	0.96	0.97	1.56	(-0.02 to 0.01)	0.12
Anxiety (4-item) - Depression (4-item)	0.62	0.61	0.17	(-0.10 to 0.13)	0.86
Anxiety (4-item) - Depression (2-item)	0.63	0.59	0.68	(-0.08 to 0.16)	0.59
Anxiety (4-item) – Anxiety (2-item)	0.92	0.90	1.25	(-0.01 to 0.05)	0.21

Table 8.

EPDS case classification at three months, six months and combined observation points as a function of established threshold criteria.

EPDS	N items	Time	Threshold				
			9/10	12/13	3/4	4/5	7/8
Cox et al. (1987)	10	All data	336(73)/126(27)	397(86)/65(14)			
Gollan et al. (2017)	7	All data			278(60)/184(40)	322(70)/140(30)	411(89)/51(11)
Cox et al. (1987)	10	3 months	189(77)/56(23)	219(89)/26(11)			
Gollan et al. (2017)	7	3 months			155(63)/90(37)	175(71)/70(29)	226(92)/19(8)
Cox et al. (1987)	10	6 months	147(68)/70(32)	178(82)/39(18)			
Gollan et al. (2017)	7	6 months			123(57)/94(43)	147(68)/70(32)	185(85)/32(15)

Note: Gollan et al.'s (2017) EPDS-7 thresholds were determined to be equivalent to standard 10-item EPDS thresholds by use of receiver operating chara cteristic analysis, thus EPDS-7 3/4 = EPDS 9/10 and thus EPDS-7 4/5 = EPDS 12/13. The EPDS-7 threshold of 7/8 is included as Gollan et al. (2017) estimated this to be equivalent to a higher EPDS threshold of 13/14.