This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of reproductive and infant psychology on 22 Dec 2020, available online: https://doi.org/10.1080/02646838.2020.1861226

Fear of childbirth measurement: Appraisal of the content overlap of four instruments

Colin R. Martin[a,*], Catriona Jones[a], Claire A. Marshall[b], Chao Huang[a], Joanne Reeve[c], Mick P. Fleming[d], Julia König[e] and Julie Jomeen[f]

[a]Institute for Clinical and Applied Health Research (ICAHR), University of Hull, Hull, UK; [b]East Yorkshire Perinatal Mental Health Liaison Team, Humber Teaching NHS Foundation Trust, Hull, UK; [c]Academy of Primary Care, Hull York Medical School, University of Hull, Hull, UK; [d]Faculty of Wellbeing, University College Isle of Man, Isle of Man, UK; [e]Department of Clinical and Biological Psychology, Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany; [f]Southern Cross University, Lismore, New South Wales, Australia.

Correspondence*

Professor Colin R. Martin

Professor of Perinatal Mental Health | Faculty of Health Sciences

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

Fear of childbirth measurement: Appraisal of the content overlap of four instruments

Abstract

Objective: To evaluate empirically the degree of content overlap between four self-report measures of fear of childbirth (FoC) identified as 'best in class' by a recent review.

Background: FoC and tokophobia is an area of increasing clinical concern and has been linked to poor maternal and neonatal outcomes. Clinical pathways have been established to improve care and interventions for FoC however, ambiguity and inconsistency remain regarding the most appropriate assessment measures.

Method: A multi-rater and consensus content analysis was undertaken to determine the degree of overlap between four 'best in class' measures of FoC/tokophobia.

Results: The Slade-Pais expectations of childbirth scale (SPECS) was found to be the preferred measure in terms of symptom overlap of the tools evaluated, however, the overall level of overlap among these measures was weak.

Conclusion: Limitations inherent to the current battery of preferred measures of FoC suggests both the desirability and urgency to develop a theoretically-grounded, psychometrically robust and accurate FoC assessment measure. Current measures of FoC are not interchangeable.

Introduction

Fear of childbirth (FoC) represents an area of contemporary clinical concern and research focus and describes anxiety related to the event of childbirth (Dencker et al., 2019; Nilsson et al., 2018). The underlying complexity of FoC as a phenomenon is important to unpack both within the context of identification (Richens, Smith, & Lavender, 2018; Toohill, Creedy, Gamble, & Fenwick, 2015) and intervention (Fenwick et al., 2015; Klabbers, Wijma, Paarlberg, Emons, & Vingerhoets, 2019; Toohill et al., 2014). The urgency for this is not only in relation to the well-being of the woman herself, but also in relation to the potential impact on the future relationship between mother and baby (Pazzagli et al., 2015). An existing tension in the literature relates to differentiating (or not) FoC from tokophobia (tocophobia), the challenge being that FoC may be conceptualised within a continuum model where a degree of FoC may be anticipated to be both normal and anticipated (Richens et al., 2018). Tokophobia implies *severe* FoC and thus a dichotomy between absence/presence of a clinical presentation that is defined as a specific phobia (Hofberg & Brockington, 2000; Striebich, Mattern, & Ayerle, 2018). FoC and tokophobia appear to be used interchangeably within the literature, a context of equipoise with doubtful support given fundamental disagreement between the notion of childbirth-related fear as a continuum (Richens et al., 2018) or distinct pathological state (Poggi, Goutaudier, Sejourne, & Chabrol, 2018). A definitive position on the continuum/state differentiation is unlikely to be realised soon, particularly given these distinctions represent seemingly irreconcilable positions in many areas of mental health, for example schizophrenia (Fleming & Martin, 2011, 2012). The recent development of clinical pathways specific to FoC (Jones, Marshall, Martin, & Jomeen, 2020) represent a step-change in access to, and provision of, evidence-based interventions in severe FoC or tokophobia such as

cognitive-behavioural therapy (Larsson et al., 2017). Somewhat surprisingly, the most fundamental component of a clinical referral pathway, the initial screen, remains an area of ambiguity in terms of choice of tool (Pallant et al., 2016; Richens et al., 2018) and associated confidence in the measurement acuity of such tools that are available (Konig, 2019).

A number of measurement tools appropriate for FoC screening have been developed ranging from single items to questionnaires with 70+ items (Richens et al., 2018). The most widely-used measure to date has been the 33-item Wijma Delivery Expectancy Questionnaire (W-DEQ; Wijma, Wijma, & Zar, 1998), in many respects, the 'gold standard' of the FoC screening genre due to both extensive use in clinical research (Nilsson et al., 2018) and wide-spread translation and validation internationally (Korukcu, Kukulu, & Firat, 2012; MoghaddamHosseini et al., 2019; Mortazavi, 2017; Takegata et al., 2013). A central tenet of the W-DEQ and indeed a core feature of both its use and conceptual underpinning is that the tool assesses a single dimension of FoC (Wijma et al., 1998), thus interpretation is based on the total score and preferred screening threshold level (Nilsson et al., 2018). This core ascribed attribute (uni-dimensionality), presents a significant challenge to the conceptual alignment of the W-DEQ to clinical application in that, invariably, studies that have examined the underlying measurement characteristics of the W-DEQ using factor analysis have found the tool to indeed be multi-dimensional (Fenaroli et al., 2019; Johnson & Slade, 2002; Konig, 2019; Pallant et al., 2016). Challenges to the accepted dimensionality of a screening measure may foster new and useful insights in what a tool really does measure. The Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983) for example, is conceptualised and scored as a two-dimensional (anxiety and depression) measure but has been found to be tri-dimensional in many studies, for

example (Christensen et al., 2020), findings that has been valuable in contextualising the measure within an alternative and coherent model of depression (Martin, Thompson, & Barth, 2008). However, such insights are generally based on consistency between factor analytic findings across studies. Many of these are specific to the W-DEQ. Findings from factor analysis studies vary widely, with large variation between the number of factors found and fundamental differences in the pattern of item-factor loadings observed (Fenaroli et al., 2019; Johnson & Slade, 2002; Mortazavi, 2017; Pallant et al., 2016). It has been suggested that translations of the W-DEQ may be affected significantly by cultural context and the translation of some items may be problematic (Richens et al., 2018). However the relative merits of this perspective are profoundly limited by the measurement model of the W-DEQ being uni-dimensional (Wijma et al., 1998). The most extensive measurement model evaluation of the W-DEQ was undertaken by Pallant et al. (2016) using both exploratory and confirmatory factor analysis and in addition, a Rasch analysis for scale and sub-scale uni-dimensionality. Pallant et al. (2016) and colleagues concluded from their analysis that (a) the W-DEQ is multi-dimensional comprising four distinct factors, (b) a shortened revision may be useful with redundant items removed and, (c) the W-DEQ should not be used in its current form.

Given the above concerns about the W-DEQ and the length of the tool for practical clinical application (Richens et al., 2018), a short instrument, such as the Fear of Birth Scale (FOBS; H. Haines, Pallant, Karlstrom, & Hildingsson, 2011), may have greater potential, particularly within the context of both research *and* clinical practice (Richens, Campbell, & Lavender, 2019). The FOBS does seem promising for an initial screen, comprising just two items (fear and worry) both scored on a 10 cm visual analogue scale and a mean score taken to compare against threshold. Circumscribed by brevity,

the FOBS has been shown to be as effective as the W-DEQ for screening (Richens et al., 2018) and for the pragmatics of stepped screening within a clinical pathway, the notion of using both tools has garnered interest (Jones et al., 2020). A concern raised regarding the W-DEQ has been inconsistent threshold/cut-off scores for the identification of severe or significant FoC, a concern that may also be inferred from studies of the FOBS which have indicated varying cut-off scores (H. M. Haines et al., 2015; Ternstrom, Hildingsson, Haines, & Rubertsson, 2015) and even utilised alternative thresholds within the same study (Richens et al., 2019). A recent investigation also highlighted what may be a fundamental limitation of the FOBS, namely inherent measurement error, the suggestion being that at the very least the FOBS requires further psychometric appraisal and potentially modification (Richens et al., 2019).

The literature thus presents a service provision conundrum, the goal of providing evidence-based support for clinically-relevant FoC against lack of an agreed definition (Nilsson et al., 2018) and significant limitations in screening measures in relation to the two most widely-used tools (Pallant et al., 2016; Richens et al., 2019; Sheen, Slade, Balling, & Houghton, 2018). Consequently, the operationalising of a clinical pathway in these circumstances is limited not only in terms of accurate screen and thus appropriate access to a service but also in relation to accurate assessment of outcomes, since both the W-DEQ and FOBS are also used to assess intervention efficacy (Jones et al., 2020).

Recognising that the precipitant of the issues above are largely a function of a lack of an unambiguous and evidence-based construct of FoC, Sheen and Slade (2018) undertook a meta-synthesis of the literature to identify and understand the content of FoC from the

woman's perspective. The findings from this work was incorporated into a further study, combined with in-depth interviews with pregnant women experiencing FoC and midwives to identify the underlying components of FoC which may be used to develop measurement tools that are theoretically and conceptually anchored (Slade, Balling, Sheen, & Houghton, 2019; Slade, Pais, Fairlie, Simpson, & Sheen, 2016). Slade et al. (2019) detailed the next stage in their stepwise project was to examine women's appraisal of items used in existing measures and mapping the constructs from their study on to these tools. A recently published study (Sheen et al., 2018; Slade, Balling, Sheen, & Houghton, 2020) highlighted the potential use of four instruments, these being the W-DEQ, the FOBS, the Slade-Pais Expectations of Childbirth Scale (SPECS; Slade et al., 2016) and the Oxford Worries about Labour Scale (OWLS; Redshaw, Martin, Rowe, & Hockley, 2009). The SPECS is a 50-item multi-dimensional measure of birth expectancy of which fear of childbirth represents a distinct sub-scale (10-items) as well as many items which are conceptually related to fear, for example loss of control. Twenty-six items from the SPECS have been suggested to be used in clinical practice for identification of FoC, though as far as the authors are aware this measure is currently in clinical use at one site in the UK[1].

Uniquely among the four instruments, the OWLS was never conceived as an instrument to assess any domain of FoC. Indeed, the OWLS was developed and originally validated as a nine-item multi-dimensional measure of worries about the labour experience. The OWLS has been or is planned to be used in a number of studies none of which has a primary focus on FoC (Henderson, Jomeen, & Redshaw, 2018; Henderson & Redshaw, 2016; Krusche, Crane, & Dymond, 2019; Roch et al., 2018). The OWLS assess three distinct but correlated domains of distress, uncertainty and

[1] The use of the 26-items from the SPECS for FoC assessment and the use of the measure in one site in the UK comes from personal communication with the SPECS study lead author.

interventions with respect to labour. Despite profound conceptual heritage and measurement characteristic differences between these four measures, it is important to be aware that the selection of these measures as representing key aspects of relevance of FoC to women themselves is representative of an endpoint of exhaustive reviews of the literature (Sheen & Slade, 2018; Sheen et al., 2018; Slade et al., 2020) and in-depth interviews with practitioners and women experiencing FoC (Slade et al., 2019). Further, these measures and in particular, the SPECS and the OWLS received endorsement from women themselves as including items that best represents their experience (Sheen et al., 2018).

Reflecting on the range of instruments that are used to assess FoC, the observation that they are used interchangeably without a selection rationale and the context that conceptually FoC itself has only recently received focused attention in conceptual alignment from the woman's perspective (Sheen & Slade, 2018; Slade et al., 2019). The overlap between the four measures highlighted by Sheen et al. (2018) and Slade et al. (2020) is of interest to appraise for two reasons. Firstly, the rich qualitative insights that have led to a focus on these four tools has yet to be triangulated using a quantitative approach. This is methodologically relevant because limitations in existing tools has highlighted the quantitative aspects of measurement to a significant degree, for example the work of Sheen and Slade (2018) and Slade et al. (2019). Secondly, the notion that measures of a concept may be used interchangeably has been emphasised as a highly contentious practice (Fried, 2017). Indeed, such assumed equivalence of measures may be one of the contributors to the 'replicability crisis' currently confronting the behavioural sciences (Anderson & Maxwell, 2017; Bardi & Zentner, 2017; Coiera, Ammenwerth, Georgiou, & Magrabi, 2018; Loken & Gelman, 2017). The findings from Sheen et al. (2018) and Slade et al. (2020) regarding the use of the FOBS,

W-DEQ, SPECS and OWLS in terms of representing, to a lesser or greater degree, women's symptoms and experience may suggest that these measures could be used interchangeably if there is sufficient overlap in symptoms within the scales. Recent influential work in this area by Fried (2017) in relation to depression, where self-report measures are indeed used interchangeably, has found that the degree of overlap between measures to be low and that this may be a significant contributor to replicability failure issues. Given the diversity inherent in the measures of FoC outlined above, Fried's (2017) perspective would appear relevant to investigate in the context of these tools.

Aim

The aim of the current investigation was to evaluate the four instruments identified by Sheen et al. (2018) and Slade et al. (2020) as best representing women's experience of FoC in terms of overlap of symptoms intrinsic within each scale across scales.

Methods

Using an adaptation of the approach of Fried (2017), an empirical content analysis of the FOBS, W-DEQ, SPECS and OWLS was undertaken to evaluate item overlap across the scales. The approach of Fried (2017) was modified for two key reasons. Firstly, Fried (2017) evaluated depression screening measures against an established diagnostic entity, specifically symptoms associated with depression, in terms of guiding the selection of the items from each questionnaire in that they may be meaningfully compared. Consequently, in that study from a total of 125 items in the seven measures evaluated, less than half of the items were evaluated for overlap by condensing items to specific depression symptoms. However, in the case of FoC, a comparable diagnostic entity does not currently exist, therefore, across the four measures all items were included for overlap comparison. Statistically exquisite and undoubtedly

methodologically innovative as Fried's (2017) study was, the author himself highlighted that the condensation of items for comparison was subjective. Fried (2017) also empahsised that a less conservative approach would be to consider all items. Reflecting on the above our approach was therefore to select all 70 items from the four measures for overlap analysis to ensure all FoC experiences that are captured by the tools are incorporated into the analysis without subjective bias. Fried (2017) differentiated between specific symptoms, those that would be generally considered more or less identical between items and compound symptoms, those which shared a high degree of similarity between items but were not equal. We adopted the same approach, but differentiated between high overlap and moderate/modest overlap for our overlap categorisation, all compared to a no overlap categorisation. Secondly, a further elaboration of Fried (2017) methodology to the current study was that we conducted four content analyses (in contrast to a single content analysis in Fried (2017)) based on the appraisal of overlap by practitioners knowledgeable in the area of FoC and who had developed a clinical pathway for FoC[2]. Three further content analyses was also conducted by academic and clinical colleagues with no significant specific knowledge of FoC but with familiarity with content analysis, thus offering an opportunity to evaluation any marked variability between those with and without specialist FoC knowledge. A consensus content analysis was then constructed by using the modal value (mode score) across raters for each overlap score. The consensus content analysis was then subject to statistical analysis.

Statistical analysis

[2] Personal communication with Dr Fried supported the adoption of multiple content analyses to offer enhanced rigour in terms of inter-rater reliability.

Consistent with Fried (2017), the content overlap of items was estimated using the Jaccard Index (JI). This metric represents a similarity coefficient specifically for binary data with a 0-1 range where 0 represents an absence of overlap and 1 represents complete overlap. The calculation of the JI is described in detail in Fried (2017) and within the context of the current investigation binary classification for calculating the JI is by collapsing high and moderate overlap classification into one category for comparison against no overlap classification, thus a dichotomous categorisation. We also adopted the same criterion used by Fried (2017) for evaluation of the strength of the JI correlation coefficient. The criteria of Evans (1996) ranges from very weak (0.00-0.19), increasing incrementally through weak, moderate, strong to very strong (0.80-1.00). Consistent with Fried (2017) and as an abstraction of the specific/compound conceptualisation of scale overlap we calculated separate item high vs. moderate overlap estimations across scales and also calculated the frequency of idiosyncratic items per scale, essentially those that appeared in no other scale.

Calculation of the JI was undertaken by the construction of a binary matrix and then all data analysis was conducted using the R programming language, version 3.6.2 (R Core Team, 2019) and the R programs, qgraph (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012), ggplot2 (Wickham, 2009), data.table (Dowle & Srinivasan, 2019), reshape2 (Wickham, 2007), psych (Revelle, 2019), ade4 (Dray & Dufour, 2007), boot (Canty & Ripley, 2019; Davison & Hinkley, 1997), irr (Gamer, Lemon , Fellows, & Singh, 2019) and viridis (Garnier, 2018). The R code for the analysis was adapted from that of Fried (2017) and incorporated an amendment to the original code (Fried, 2020).

Inter-rater reliability was calculated across the seven raters for each combination of responses using Fleiss' kappa (Fleiss, 1971) with level of agreement determined by reference to the thresholds of Landis and Koch (1977).

Results

Seven content analyses were completed individually by four specialist practitioners in perinatal mental health (drawn from the disciplines of nursing, midwifery and psychology), a mental health nursing practitioner, a medical general practitioner and a statistician. Fleiss' kappa calculated for each instrument revealed significant agreement between raters (Table 1.) ranging from fair to moderate agreement with reference to the criteria of Landis and Koch (1977).

TABLE 1. ABOUT HERE

The JI overlap index for each scale by each rater is summarised in Table 2. Notwithstanding variability between raters, the SPECS was found to have consistently the most overlap. The consensus content analysis is also summarised in Table 2. revealing the SPECS to have the most overlap and the OWLS the least.

TABLE 2. ABOUT HERE

The JI correlation coefficients for the consensus content analysis are summarised in Table 3. The mean JI index across scales based on the consensus content analysis was 0.247 which according to the criteria of Evans (1996) is a weak level of overlap.

TABLE 3. ABOUT HERE

The OWLS and the FOBS were observed to capture the lowest percentage similarity of the total seventy items at 26% (18 items) each, while the SPECS and W-DEQ captured the most at 67% (47 items) each. The total number of idiosyncratic items, those not represented in any other scale was 29 (41%). The FOBS had no idiosyncratic items, whereas the SPECS had 7 (27%), the W-DEQ had 16 (48%) and the OWLS had 6 (67%).

Comparison between items across scales in relation to degree of similarity (high vs. moderate overlap vs. no overlap) is summarised in Figure 1.

FIGURE 1. ABOUT HERE

Discussion

The findings from the current investigation raise a number of questions concerning the measurement of FoC. The weak level of overlap between measures is highly indicative that the tools are not interchangeable, thus confirming a significant inherent source of error if comparisons are made between studies based on different FoC measures. These findings are thus consistent with the assertion of Fried (2017) that assumed interchangeability of measures is erroneous. This is fundamentally important because of the suggestion that this is a potential contributory factor to the replicability crisis (Fried, 2017). Moreover, these findings are consistent with Sheen et al. (2018) and Slade et al. (2020) which revealed, from women's perspectives, that measures of FoC symptoms of are not comprehensively captured by one particular instrument. However, the current study does indicate that in terms of overlap across scales, as assessed by the

JI, the SPECS would seem to be the current tool of preference in terms of overlap across scales, thus capturing the larger component of symptoms across the total seventy items of all scales combined. It is noteworthy that though the W-DEQ captured an identical *absolute* percentage of symptoms across scales as the SPECS (67%), the W-DEQ is also a longer measure and more importantly had a much larger percentage of idiosyncratic items (48% vs. 27%) than the SPECS. Further supportive evidence for the preference for the SPECS can be inferred from the individual content analyses, which though exhibiting a degree of variability between raters, also consistently found the SPECS to offer most overlap across tools.

The caveat in suggesting the SPECS is the preferred measure from the four evaluated in the current study must be the weak level of overlap between scales which suggests the development of an experience-informed, theoretically-grounded and psychometrically robust measure of FoC is a pressing contemporary need, particularly given the pre-eminence of accurate assessment within the establishment of clinical pathways (Jones et al., 2020). The OWLS in contrast to the SPECS had very little overlap and the highest percentage of idiosyncratic items. This perhaps should not be an entirely surprising finding as uniquely among the four tools it was never designed to be a measure of FoC. However, it is also important to reflect that the OWLS was selected by a review of measures and evaluated by women with FoC to be a measure which represented their experiences (Sheen et al., 2018). The SPECS in contrast was designed to intrinsically assess FoC and followed a robust instrument development process (Slade et al., 2016), however, the findings from Sheen et al. (2018) and Slade et al. (2020) in terms of the diversity of the four tools and women's experiences would indicate that the SPECS does not assess the core aspects of FoC *comprehensively*. However, as

highlighted by (Slade et al., 2020), none of the four instruments were optimal in terms of content validity, understanding and acceptability from the woman's perspective.

The study had one important limitation. The approach to content analysis taken is relatively novel and therefore findings must be tempered within the parameters of an approach which has yet to penetrate the mainstream literature. Further, although the statistical analysis undertaken was sophisticated, the use of content analysis and multiple raters is an established approach and it is hoped also contextually sensitive to the qualitative research (Sheen & Slade, 2018; Sheen et al., 2018; Slade et al., 2019, 2020) which underpinned the approach taken for the current study.

Given that even the SPECS was observed to have inherent limitations in assessing all key aspects of FoC, future research to develop a definitive measure of FoC that addresses these deficits is suggested. This may meaningfully incorporate the elements of the measures in the current study that both overlap and are appraised by women to be representative and sensitive to their individual experience. Further, since the approach to content analysis undertaken was found to be both useful and insightful, application to other aspects of perinatal mental health, such as anxiety, and perinatal wellbeing, such as quality of life, is also suggested.

Finally, by highlighting issues related to tools that have been either specifically designed to or suggested could be used to assess FoC, it is useful to consider that a further crucial need is to develop an evidence-based and universally agreed definition of FoC from which measures can be conceptually grounded (Jomeen et al., 2020). Indeed, the recently published consensus statement by Jomeen and colleagues emphasises the

potentially negative clinical implications of the current rudimentary theoretical and knowledge base regarding FoC. Central to this is the impact on adequacy of screening, assessment and intervention and these have been emphasised as key areas of pressing future research and highlighted within the consensus statement is the use of measures and an understanding of their implicit measurement characteristics (Jomeen et al., 2020).

In summary, the current study took a methodologically novel approach to reflect upon and consider a fundamental component of the FoC literature, namely the accurate and appropriate measurement of the concept by existing measures. Principally informed by qualitative research, our quantitative approach has indicated not only a preference for the SPECS among the instruments evaluated but also highlighted the limitations of the same.

Acknowledgments

The first author is extremely grateful to Dr Eiko Fried, Department of Clinical Psychology, Leiden University, for helpful advice and suggestions regarding the statistical analysis and discussion around statistical coding and code modification and also to both Dr Fried and Dr Jana Jarecki, Department of Economic Psychology, University of Basel for the original base coding that was used for Figure 1. We are grateful also for the expert opinion and advice of two anonymous reviewers of the manuscript.

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

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

Canty, A., & Ripley, B. (2019). boot: Bootstrap R (S-Plus) Functions. R package (Version 1.3-23.).

Christensen, A. V., Dixon, J. K., Juel, K., Ekholm, O., Rasmussen, T. B., Borregaard, B., . . . Berg, S. K. (2020). Psychometric properties of the Danish Hospital Anxiety and Depression Scale in patients with cardiac disease: results from the DenHeart survey. *Health and Quality of Life Outcomes, 18*(1), 9. doi:10.1186/s12955-019-1264-0

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

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.

Dencker, A., Nilsson, C., Begley, C., Jangsten, E., Mollberg, M., Patel, H., . . . Sparud-Lundin, C. (2019). Causes and outcomes in studies of fear of childbirth: A systematic review. *Women and Birth, 32*(2), 99-111. doi:10.1016/j.wombi.2018.07.004

Dowle, M., & Srinivasan, A. (2019). data.table: Extension of `data.frame' (Version 1.12.8). Retrieved from https://CRAN.R-project.org/package=data.table

Dray, S., & Dufour, A. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software, 22*, 1-20. doi:10.18637/jss.v022.i04

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software, 48*. Retrieved from http://www.jstatsoft.org/v48/i04/.

Evans, J. D. (1996). *Straightforward Statistics for the Behavioural Sciences*. Pacific Grove, CA: Brooks/Cole publishing.

Fenaroli, V., Molgora, S., Dodaro, S., Svelato, A., Gesi, L., Molidoro, G., . . . Ragusa, A. (2019). The childbirth experience: obstetric and psychological predictors in Italian primiparous women. *BMC Pregnancy and Childbirth, 19*(1), 419. doi:10.1186/s12884-019-2561-7

Fenwick, J., Toohill, J., Gamble, J., Creedy, D. K., Buist, A., Turkstra, E., . . . Ryding, E. L. (2015). Effects of a midwife psycho-education intervention to reduce childbirth fear on women's birth outcomes and postpartum psychological wellbeing. *BMC Pregnancy and Childbirth, 15*, 284. doi:10.1186/s12884-015-0721-y

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fleming, M. P., & Martin, C. R. (2011). Genes and schizophrenia: a pseudoscientific disenfranchisement of the individual. *Journal of Psychiatric and Mental Health Nursing, 18*(6), 469-478. doi:10.1111/j.1365-2850.2011.01690.x

Fleming, M. P., & Martin, C. R. (2012). From classical psychodynamics to evidence synthesis: the motif of repression and a contemporary understanding of a key mediatory mechanism in psychosis. *Current Psychiatry Reports, 14*(3), 252-258. doi:10.1007/s11920-012-0260-4

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders, 208*, 191-197. doi:10.1016/j.jad.2016.10.019

Fried, E. I. (2020). Corrigendum to "The 52 symptoms of major depression: lack of content overlap among seven common depression scales", [Journal of Affective Disorders, 208, 191-197]. *Journal of Affective Disorders, 260*, 744. doi:10.1016/j.jad.2019.05.029

Gamer, M., Lemon , J., Fellows, I., & Singh, P. (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package (Version 0.84.1.). Retrieved from https://CRAN.R-project.org/package=irr

Garnier, S. (2018). viridis: Default Color Maps from 'matplotlib' (Version 0.5.1). Retrieved from https://CRAN.R-project.org/package=viridis

Haines, H., Pallant, J. F., Karlstrom, A., & Hildingsson, I. (2011). Cross-cultural comparison of levels of childbirth-related fear in an Australian and Swedish sample. *Midwifery, 27*(4), 560-567. doi:10.1016/j.midw.2010.05.004

Haines, H. M., Pallant, J. F., Fenwick, J., Gamble, J., Creedy, D. K., Toohill, J., & Hildingsson, I. (2015). Identifying women who are afraid of giving birth: A comparison of the fear of birth scale with the WDEQ-A in a large Australian cohort. *Sex Reprod Healthc, 6*(4), 204-210. doi:10.1016/j.srhc.2015.05.002

Henderson, J., Jomeen, J., & Redshaw, M. (2018). Care and self-reported outcomes of care experienced by women with mental health problems in pregnancy: Findings

from a national survey. *Midwifery, 56*, 171-178. doi:10.1016/j.midw.2017.10.020

Henderson, J., & Redshaw, M. (2016). Worries About Labor and Birth: A Population-Based Study of Outcomes for Young Primiparous Women. *Birth, 43*(2), 151-158. doi:10.1111/birt.12219

Hofberg, K., & Brockington, I. (2000). Tokophobia: an unreasoning dread of childbirth. A series of 26 cases. *British Journal of Psychiatry, 176*, 83-85. doi:10.1192/bjp.176.1.83

Johnson, R., & Slade, P. (2002). Does fear of childbirth during pregnancy predict emergency caesarean section? *BJOG: An International Journal of Obstetrics and Gynaecology, 109*(11), 1213-1221. doi:10.1046/j.1471-0528.2002.01351.x

Jomeen, J., Martin, C. R., Jones, C., Marshall, C., Ayers, S., Burt, K., . . . Thomson, G. (2020). Tokophobia and fear of birth: A workshop consensus statement on current issues and recommendations for future research. *Journal of Reproductive and Infant Psychology*. doi:10.1080/02646838.2020.1843908

Jones, C., Marshall, C., Martin, C. R., & Jomeen, J. (2020). Pregnant with fear. *Community Practitioner, 2*. Retrieved from https://www.communitypractitioner.co.uk/features/2020/02/pregnant-fear

Klabbers, G. A., Wijma, K., Paarlberg, K. M., Emons, W. H. M., & Vingerhoets, A. (2019). Haptotherapy as a new intervention for treating fear of childbirth: a randomized controlled trial. *Journal of Psychosomatic Obstetrics and Gynaecology, 40*(1), 38-47. doi:10.1080/0167482X.2017.1398230

Konig, J. (2019). The German W-DEQ version B-Factor structure and prediction of posttraumatic stress symptoms six weeks and one year after childbirth. *Health*

Care for Women International, 40(5), 581-596. doi:10.1080/07399332.2019.1583230

Korukcu, O., Kukulu, K., & Firat, M. Z. (2012). The reliability and validity of the Turkish version of the Wijma Delivery Expectancy/Experience Questionnaire (W-DEQ) with pregnant women. *Journal of Psychiatric and Mental Health Nursing, 19*(3), 193-202. doi:10.1111/j.1365-2850.2011.01694.x

Krusche, A., Crane, C., & Dymond, M. (2019). An investigation of dispositional mindfulness and mood during pregnancy. *BMC Pregnancy and Childbirth, 19*(1), 273. doi:10.1186/s12884-019-2416-2

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/843571

Larsson, B., Karlstrom, A., Rubertsson, C., Ternstrom, E., Ekdahl, J., Segebladh, B., & Hildingsson, I. (2017). Birth preference in women undergoing treatment for childbirth fear: A randomised controlled trial. *Women and Birth, 30*(6), 460-467. doi:10.1016/j.wombi.2017.04.004

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584-585. doi:10.1126/science.aal3618

Martin, C. R., Thompson, D. R., & Barth, J. (2008). Factor structure of the Hospital Anxiety and Depression Scale in coronary heart disease patients in three countries. *Journal of Evaluation in Clinical Practice, 14*(2), 281-287. doi:10.1111/j.1365-2753.2007.00850.x

MoghaddamHosseini, V., Makai, A., Varga, K., Acs, P., Premusz, V., & Varnagy, A. (2019). Assessing fear of childbirth and its predictors among Hungarian pregnant women using Wijma Delivery Expectancy/Experience Questionnaire

subscales. *Psychology, Health & Medicine, 24*(7), 879-889. doi:10.1080/13548506.2019.1572904

Mortazavi, F. (2017). Validity and reliability of the Farsi version of Wijma delivery expectancy questionnaire: an exploratory and confirmatory factor analysis. *Electronic Physician, 9*(6), 4606-4615. doi:10.19082/4606

Nilsson, C., Hessman, E., Sjoblom, H., Dencker, A., Jangsten, E., Mollberg, M., . . . Begley, C. (2018). Definitions, measurements and prevalence of fear of childbirth: a systematic review. *BMC Pregnancy and Childbirth, 18*(1), 28. doi:10.1186/s12884-018-1659-7

Pallant, J. F., Haines, H. M., Green, P., Toohill, J., Gamble, J., Creedy, D. K., & Fenwick, J. (2016). Assessment of the dimensionality of the Wijma delivery expectancy/experience questionnaire using factor analysis and Rasch analysis. *BMC Pregnancy and Childbirth, 16*(1), 361. doi:10.1186/s12884-016-1157-8

Pazzagli, C., Laghezza, L., Capurso, M., Sommella, C., Lelli, F., & Mazzeschi, C. (2015). Antecedents and consequences of fear of childbirth in nulliparous and parous women. *Infant Ment Health Journal, 36*(1), 62-74. doi:10.1002/imhj.21483

Poggi, L., Goutaudier, N., Sejourne, N., & Chabrol, H. (2018). When Fear of Childbirth is Pathological: The Fear Continuum. *Maternal and Child Health Journal, 22*(5), 772-778. doi:10.1007/s10995-018-2447-8

R Core Team. (2019). R: A language and environment for statistical computing. . Vienna, Austria.: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Redshaw, M., Martin, C., Rowe, R., & Hockley, C. (2009). The Oxford Worries about Labour Scale: women's experience and measurement characteristics of a

measure of maternal concern about labour and birth. *Psychology, Health & Medicine, 14*(3), 354-366. doi:10.1080/13548500802707159

Revelle, W. (2019). psych: Procedures for Personality and Psychological Research (Version 1.9.12). Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Richens, Y., Campbell, M., & Lavender, T. (2019). Fear of birth-A prospective cohort study of primigravida in the UK. *Midwifery, 77*, 101-109. doi:10.1016/j.midw.2019.06.014

Richens, Y., Smith, D. M., & Lavender, D. T. (2018). Fear of birth in clinical practice: A structured review of current measurement tools. *Sexual and Reproductive Healthcare, 16*, 98-112. doi:10.1016/j.srhc.2018.02.010

Roch, G., Borges Da Silva, R., de Montigny, F., Witteman, H. O., Pierce, T., Semenic, S., . . . Gagnon, M. P. (2018). Impacts of online and group perinatal education: a mixed methods study protocol for the optimization of perinatal health services. *BMC Health Services Research, 18*(1), 382. doi:10.1186/s12913-018-3204-9

Sheen, K., & Slade, P. (2018). Examining the content and moderators of women's fears for giving birth: A meta-synthesis. *Journal of Clinical Nursing, 27*(13-14), 2523-2535. doi:10.1111/jocn.14219

Sheen, K., Slade, P., Balling, K., & Houghton, G. (2018). *An examination of clarity, acceptability and content validity of existing questionnaires for the measurement of fear of childbirth in a UK population*. Paper presented at the Society for Reproductive and Infant Psychology Annual conference, Lodz, Poland.

Slade, P., Balling, K., Sheen, K., & Houghton, G. (2019). Establishing a valid construct of fear of childbirth: findings from in-depth interviews with women and

midwives. *BMC Pregnancy and Childbirth, 19*(1), 96. doi:10.1186/s12884-019-2241-7

Slade, P., Balling, K., Sheen, K., & Houghton, G. (2020). Identifying fear of childbirth in a UK population: qualitative examination of the clarity and acceptability of existing measurement tools in a small UK sample. *BMC Pregnancy and Childbirth, 20*(1), 553. doi:10.1186/s12884-020-03249-4

Slade, P., Pais, T., Fairlie, F., Simpson, A., & Sheen, K. (2016). The development of the Slade–Pais Expectations of Childbirth Scale (SPECS). *Journal of Reproductive and Infant Psychology, 34*(5), 495-510. doi:10.1080/02646838.2016.1209300

Striebich, S., Mattern, E., & Ayerle, G. M. (2018). Support for pregnant women identified with fear of childbirth (FOC)/tokophobia - A systematic review of approaches and interventions. *Midwifery, 61*, 97-115. doi:10.1016/j.midw.2018.02.013

Takegata, M., Haruna, M., Matsuzaki, M., Shiraishi, M., Murayama, R., Okano, T., & Severinsson, E. (2013). Translation and validation of the Japanese version of the Wijma Delivery Expectancy/Experience Questionnaire version A. *Nursing & Health Sciences, 15*(3), 326-332. doi:10.1111/nhs.12036

Ternstrom, E., Hildingsson, I., Haines, H., & Rubertsson, C. (2015). Higher prevalence of childbirth related fear in foreign born pregnant women--findings from a community sample in Sweden. *Midwifery, 31*(4), 445-450. doi:10.1016/j.midw.2014.11.011

Toohill, J., Creedy, D. K., Gamble, J., & Fenwick, J. (2015). A cross-sectional study to determine utility of childbirth fear screening in maternity practice - An Australian perspective. *Women and Birth, 28*(4), 310-316. doi:10.1016/j.wombi.2015.05.002

Toohill, J., Fenwick, J., Gamble, J., Creedy, D. K., Buist, A., Turkstra, E., & Ryding, E. L. (2014). A randomized controlled trial of a psycho-education intervention by midwives in reducing childbirth fear in pregnant women. *Birth, 41*(4), 384-394. doi:10.1111/birt.12136

Wickham, H. (2007). Reshaping Data with the reshape Package. Journal of Statistical Software. *Journal of Statistical Software, 21*, 1-20. Retrieved from http://www.jstatsoft.org/v21/i12/

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.

Wijma, K., Wijma, B., & Zar, M. (1998). Psychometric aspects of the W-DEQ; a new questionnaire for the measurement of fear of childbirth. *Journal of Psychosomatic Obstetrics and Gynaecology, 19*(2), 84-97. doi:10.3109/01674829809048501

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica, 67*(6), 361-370. doi:10.1111/j.1600-0447.1983.tb09716.x

Table 1. Fleiss' kappa ratings for each scale at three-point and dichotomous agreement level (first order normal approximation confidence intervals calculated by bootstrap based on 10,000 replicates).

Scale	Level	Kappa	Std. error	95%(CI)	Z	*p*	Interpretation
FOBS	Three	0.275	0.051	0.179 - 0.381	14.10	<0.05	Fair
W-DEQ	Three	0.547	0.044	0.463 - 0.636	28.50	<0.05	Moderate
SPECS	Three	0.537	0.046	0.449 - 0.630	28.30	<0.05	Moderate
OWLS	Three	0.463	0.064	0.343 - 0.595	23.70	<0.05	Moderate
FOBS	Two	0.381	0.060	0.268 - 0.502	14.60	<0.05	Fair
W-DEQ	Two	0.567	0.059	0.456 - 0.688	21.70	<0.05	Moderate
SPECS	Two	0.508	0.059	0.397 - 0.629	19.50	<0.05	Moderate
OWLS	Two	0.503	0.064	0.382 - 0.634	19.30	<0.05	Moderate

Note: Z = Z-score, interpretation is based on Landis and Koch (1977) values of 0–0.20 = slight agreement, 0.
agreement, 0.41–0.60 = moderate agreement , 0.61–0.80 = substantial agreement, and 0.81–1.00 near perfect

Table 2. Jaccard Index overlap estimations for each scale by each rater and final consensus rating.

	FOBS	W-DEQ	SPECS	OWLS
Rater 1.	0.199	0.172	0.249	0.079
Rater 2.	0.417	0.443	0.508	0.417
Rater 3.	0.305	0.432	0.471	0.247
Rater 4.	0.476	0.455	0.522	0.267
Rater 5.	0.360	0.318	0.412	0.118
Rater 6.	0.489	0.527	0.617	0.499
Rater 7.	0.167	0.236	0.291	0.096
Consensus	0.247	0.281	0.350	0.111

Table 3. Jaccard Index correlation coefficients between Fear of Childbirth scales and summary of individual measure overlap for the consensus content analysis.

	FOBS	W-DEQ	SPECS	OWLS
FOBS	1	0.327	0.354	0.059
W-DEQ	0.327	1	0.469	0.048
SPECS	0.354	0.469	1	0.226
OWLS	0.059	0.048	0.226	1

Figure 1. Relationship of clinical screening measures for Fear of Childbirth at individual item level and degree of symptom similarity.

Symptom Number	Symptom content
1	Calm/worried
2	No fear/strong fear
3	Fantastic
4	Frightful
5	Lonely
6	Strong
7	Confidence
8	Afraid
9	Deserted
10	Weak
11	Safe
12	Independent
13	Desolate
14	Tense
15	Glad
16	Proud
17	Abandoned
18	Composed
19	Relaxed
20	Happy
21	Panic
22	Hopelessness
23	Longing for the child
24	Self-confidence
25	Trust
26	Pain
27	Behave badly
28	Surrender control
29	Lost control
30	Funny
31	Natural
32	Self-evident
33	Dangerous
34	Fantasies child death
35	Fantasies child injury
36	My body will fail me during labour
37	I will not be able to give birth naturally
38	I will be very worried when I am in labour
39	Labour will be scary
40	Labour is unknown
41	Labour will be complicated
42	I worry I will need emergency surgery
43	I will be worried about the health of my baby
44	I will be scared
45	I will be anxious
46	I worry that labour will be extremely painful
47	I worry about the length of my labour (either too long or too short)
48	I will not be able to cope with the pain
49	I will need medication to manage the labour pain
50	I will not get the pain relief I want
51	I am emotionally strong enough to cope with labour
52	I will be hysterical during labour (during labour removed on form)
53	I will feel extremely anxious when in labour
54	I will be very worried when I am in labour
55	I worry I will lose control during labour
56	I worry I will embarrass myself
57	I will feel physically exposed during labour
58	I will be too tired to appreciate the birth
59	I will feel calm during labour
60	I worry about trauma to my body
61	My body will be hurt during labour
62	Needing a caesarean
63	Getting effective pain relief
64	Having a forceps or ventouse delivery
65	Getting to the hospital in time
66	Having to be induced
67	Having a long labour
68	Not knowing how long labour would take
69	Not knowing when I would go into labour
70	Pain and discomfort of labour