Shortening the Edinburgh Postnatal Depression Scale using Optimal Test Assembly Methods: Development of the EPDS-Dep-5

Running Title: Shortening the EPDS

Daphna Harel, PhD[1,2]; Brooke Levis, PhD[3-5]; Miyabi Ishihara, MSc[6]; Alexander W. Levis, MSc[7]; Simone N. Vigod, MD[8]; Louise M. Howard, PhD[9]; Brett D. Thombs, PhD[3,4,10-14]; Andrea Benedetti, PhD[4,10,15]; and the DEPRESSion Screening Data (DEPRESSD) EPDS Collaboration

[1]PRIISM Applied Statistics Center, New York University, New York, New York, USA; [2]Department of Applied Statistics, Social Science, and Humanities, New York University, New York, New York, USA; [3]Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; [4]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada; [5]Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire, UK; [6]Department of Statistics, University of California Berkeley, Berkeley, California, USA; [7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA; [8]Women's College Hospital and Research Institute, University of Toronto, Toronto, Ontario, Canada; [9]Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; [10]Department of Medicine, McGill University, Montréal, Québec, Canada; [11]Department of Psychiatry, McGill University, Montréal, Québec, Canada; [12]Department of Psychology, McGill University, Montréal, Québec, Canada; [13]Department of Educational and Counselling Psychology, McGill University, Montréal, Québec, Canada; [14]Biomedical Ethics Unit, McGill

University, Montréal, Québec, Canada; [15]Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, Québec, Canada

DEPRESSD EPDS Collaboration: Ying Sun, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Chen He, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Ankur Krishnan, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Yin Wu, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Parash Mani Bhandari, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Dipika Neupane, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Zelalem Negeri, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Mahrukh Imran, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Danielle B. Rice, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Marleine Azar, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Matthew J. Chiovitti, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Nazanin Saadat, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Kira E. Riehm, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Jill T. Boruff, Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montréal, Québec, Canada; Pim Cuijpers, Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health research institute, Vrije Universiteit Amsterdam, the Netherlands; Simon Gilbody, Hull York Medical School and the Department of

Health Sciences, University of York, Heslington, York, UK; John P. A. Ioannidis, Department of Medicine, Department of Health Research and Policy, Department of Biomedical Data Science, Department of Statistics, Stanford University, Stanford, California, USA; Lorie A. Kloda, Library, Concordia University, Montréal, Québec, Canada; Scott B. Patten, Departments of Community Health Sciences and Psychiatry, University of Calgary, Calgary, Canada; Ian Shrier, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Roy C. Ziegelstein, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; Liane Comeau, International Union for Health Promotion and Health Education, École de santé publique de l'Université de Montréal, Montréal, Québec, Canada; Nicholas D. Mitchell, Department of Psychiatry, University of Alberta, Edmonton, Alberta, Canada; Marcello Tonelli, Department of Medicine, University of Calgary, Calgary, Alberta, Canada; Jacqueline Barnes, Department of Psychological Sciences, Birkbeck, University of London, UK; Cheryl Tatano Beck, University of Connecticut School of Nursing, Mansfield, Connecticut, USA; Carola Bindt, Department of Child and Adolescent Psychiatry, University Medical Center Hamburg-Eppendorf, Germany; Felipe Pinheiro de Figueiredo, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, Brazil; Gracia Fellmeth, Nuffield Department of Population Health, University of Oxford, Oxford, UK; Barbara Figueiredo, School of Psychology, University of Minho, Portugal; Eric P. Green, Duke Global Health Institute, Durham, North Carolina, USA; Nadine Helle, Department of Child and Adolescent Psychiatry, University Medical Center Hamburg-Eppendorf, Germany; Pirjo A. Kettunen, Department of General Hospital Psychiatry, North Karelia Central Hospital, Joensuu, Finland; Jane Kohlhoff, School of Psychiatry, University of New South Wales, Kensington, Australia; Zoltán Kozinszky, Department of Obstetrics and Gynecology, Danderyd Hospital,

Stockholm, Sweden; Angeliki A. Leonardou, First Department of Psychiatry, Women's Mental Health Clinic, Athens University Medical School, Athens, Greece; Sandra Nakić Radoš, Department of Psychology, Catholic University of Croatia, Zagreb, Croatia; Tamsen J. Rochat, MRC/Developmental Pathways to Health Research Unit, Faculty of Health Sciences, University of Witwatersrand, South Africa; Johanne Smith-Nielsen, Center for Early intervention and Family studies, Department of Psychology, University of Copenhagen, Denmark; Alan Stein, Department of Psychiatry, University of Oxford, Oxford, UK; Robert C. Stewart, Department of Mental Health, College of Medicine, University of Malawi, Malawi; Meri Tadinac, Department of Psychology, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia; S. Darius Tandon, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA; Iva Tendais, School of Psychology, University of Minho, Portugal; Annamária Töreki, Department of Emergency, University of Szeged, Hungary; Thach D. Tran, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia; Katherine Turner, Epilepsy Center-Child Neuropsychiatry Unit, ASST Santi Paolo Carlo, San Paolo Hospital, Milan, Italy; Mette S. Væver, Centre for Early Intervention and Family Studies, Department of Psychology, University of Copenhagen, Copenhagen, Denmark; Johann M. Vega-Dienstmaier, Facultad de Medicina Alberto Hurtado, Universidad Peruana Cayetano Heredia. Lima, Perú.

Address for Correspondence:

Daphna Harel, PhD; 246 Greene Street, 3rd floor, New York NY, 10003; Tel (212) 992-6701; E-mail: daphna.harel@nyu.edu

Funding: This study was funded by the Canadian Institutes of Health Research (CIHR, KRS-140994). Dr. Levis was supported by a Fonds de recherche du Québec - Santé (FRQS) Postdoctoral Training Fellowship. Drs. Thombs and Benedetti were supported by a FRQS researcher salary awards. Dr. Wu was supported by a FRQS Postdoctoral Training Fellowship. Mr. Bhandari was supported by a studentship from the Research Institute of the McGill University Health Centre. Ms. Neupane was supported by G.R. Caverhill Fellowship from the Faculty of Medicine, McGill University. Ms. Rice was supported by a Vanier Canada Graduate Scholarship. Ms. Azar was supported by a FRQS Masters Training Award. The primary study by Barnes et al. was supported by a grant from the Health Foundation (1665/608). The primary study by Beck et al. was supported by the Patrick and Catherine Weldon Donaghue Medical Research Foundation and the University of Connecticut Research Foundation. The primary study by Helle et al. was supported by the Werner Otto Foundation, the Kroschke Foundation, and the Feindt Foundation. The primary study by de Figueiredo et al. was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo. The primary study by Tendais et al. was supported under the project POCI/SAU-ESP/56397/2004 by the Operational Program Science and Innovation 2010 (POCI 2010) of the Community Support Board III and by the European Community Fund FEDER. This primary study by Green et al. was supported by a grant from the Duke Global Health Institute (453-0751). The primary study by Kettunen et al. was supported with an Annual EVO Financing (Special govern- ment subsidies from the Ministry of Health and Welfare, Finland) by North Karelia Central Hospital and Päijät-Häme Central Hospital. The primary study by Phillips et al. was supported by a scholarship from the National Health and Medical and Research Council (NHMRC). The primary study by Nakić Radoš et al. was supported by the Croatian Ministry of Science, Education, and Sports (134-0000000-2421). The

primary study by Rochat et al. was supported by grants from the University of Oxford (HQ5035), the Tuixen Foundation (9940), the Wellcome Trust (082384/Z/07/Z and 071571), and the American Psychological Association. Dr. Rochat receives salary support from a Wellcome Trust Intermediate Fellowship (211374/Z/18/Z). The primary study by Smith-Nielsen et al. was supported by a grant from the charitable foundation Tryg Foundation (Grant ID no 107616). The primary study by Prenoveau et al. was supported by The Wellcome Trust (grant number 071571). The primary study by Stewart et al. was supported by Professor Francis Creed's Journal of Psychosomatic Research Editorship fund (BA00457) administered through University of Manchester. The primary study by Tandon et al. was funded by the Thomas Wilson Sanitarium. The primary study by Tran et al. was supported by the Myer Foundation who funded the study under its Beyond Australia scheme. Dr. Tran was supported by an early career fellowship from the Australian National Health and Medical Research Council. The primary study by Vega-Dienstmaier et al. was supported by Tejada Family Foundation, Inc, and Peruvian-American Endowment, Inc. No other authors reported funding for primary studies or for their work on the present study. No sponsor or funder was involved in the study design; in the collection, analysis and interpretation of the data; in the writing of the report; or in the decision to submit the paper for publication.

Declaration of Competing Interests: All authors have completed the ICJME uniform disclosure form and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years with the following exceptions: Dr. Vigod declares that she receives royalties from UpToDate, outside the submitted work. Dr. Tonelli declares that he has received a

grant from Merck Canada, outside the submitted work. Dr. Beck declares that she receives royalties for her Postpartum Depression Screening Scale published by Western Psychological Services. All authors declare no other relationships or activities that could appear to have influenced the submitted work. No funder had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Author Contributions: DH, BL, SNV, BDT, AB, JTB, PC, SG, JPAI, LAK, DM, SBP, IS, RCZ, LC, NDM and MTonelli were responsible for the study conception and design. JTB and LAK designed and conducted database searches to identify eligible studies. JB, CTB, CB, FPeF, GF, BF, EPG, NH, PAK, JK, ZK, AAL, SNR, TJR, JSN, AS, RCS, MTadinac, SDT, IT, AT, TDT, KTrevillion, KTurner, MSV and JMVD contributed primary datasets that were included in this study. BL, YS, CH, AK, YW, PMB, DN, ZN, MImran, DBR, MA, TAS, MJC, NS, KER and BDT contributed to data extraction and coding for the individual participant data meta-analysis. DH, MIshihara and AWL conducted analyses and interpreted results. DH and BL drafted the manuscript. All authors provided a critical review and approved the final manuscript. DH is the guarantor. BDT and AB have full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analyses.

Word Count: 3,764

Data Availability: Requests for data access should be made to Dr. Brooke Levis (brooke.levis@gmail.com).

ABSTRACT

Aims: This study used a large database to develop a reliable and valid shortened form of the Edinburgh Postnatal Depression Scale (EPDS), a self-report questionnaire used for depression screening in pregnancy and postpartum, based on objective criteria.

Methods: Item responses from the 10-item EPDS were obtained from 5,157 participants (765 major depression cases) from 22 primary screening accuracy studies that compared the EPDS to the Structured Clinical Interview for DSM (SCID). Unidimensionality of the EPDS latent construct was verified using confirmatory factor analysis, and an item response theory model was fit. Optimal test assembly (OTA) methods identified a maximally informative shortened form for each possible scale length between 1 and 9 items. The final shortened form was selected based on pre-specified validity and reliability criteria and non-inferiority of screening accuracy of the EPDS as compared to the SCID.

Results: A 5-item short form of the EPDS (EPDS-Dep-5) was selected. The EPDS-Dep-5 had a Cronbach's alpha of 0.82. Sensitivity and specificity of the EPDS-Dep-5 for a cutoff of 4 or greater were 0.83 (95% CI, 0.73, 0.89) and 0.86 (95% CI, 0.80, 0.90) and were statistically non-inferior to the EPDS. The correlation of total scores with the full EPDS was high (r = 0.91).

Conclusion: The EPDS-Dep-5 is a valid short form with minimal loss of information when compared to the full-length EPDS. The EPDS-Dep-5 was developed with OTA methods using objective, pre-specified criteria, but the approach is data-driven and exploratory. Thus, there is a need to replicate results of this study in different populations.

Keywords:

Depression; Optimal Test Assembly; Patient Reported Outcome; Short Form

Significant Outcomes:

- A 5-item short form of the EPDS can be used to screen for depression in the perinatal period.
- The 5-item short form was shown to be valid and reliable in a sample of 5,157 participants
- Optimal test assembly methods provide a replicable and reproducible methodology to shorten patient reported outcomes.

Limitations:

- This study was not able to obtain data from 25 of 81 eligible datasets.
- There exists substantial heterogeneity across studies in terms of country and language of administration of the semi-structured interview.
- The optimal test assembly procedure is data-driven and should be replicated.

INTRODUCTION

Depression is a leading cause of disability among women (Kessler, 2003). Although the 7-13% prevalence of major depression during pregnancy and postpartum (Bennett, Einarson, Taddio, Koren, & Einarson, 2004; Gavin et al., 2005; Gaynes et al., 2005; OHara & Swain, 1996) is similar to rates among women during non-childbearing periods (Cooper, Campbell, Day, Kennerley, & Bond, 1988; Cox, Murray, & Chapman, 1993; Gavin et al., 2005; Ohara, Zekoski, Philipps, & Wright, 1990; Stewart, 2011; Vesga-Lopez et al., 2008), perinatal depression is associated with adverse outcomes for the mother, developing child, mother-infant relationship and marital quality (Whitley & Kirmayer, 2008; Zelkowitz & Milet, 1996, 2001). Most women with depression in the perinatal period, however, do not receive adequate care (Duhoux, Fournier, Gauvin, & Roberge, 2013; Duhoux, Fournier, Nguyen, Roberge, & Beveridge, 2009; Howard et al., 2014). Rapidly identifying women with depression to improve their care is a high clinical priority (Canada, 2012).

The 10-item Edinburgh Postnatal Depression Scale (EPDS) is the most commonly used self-report questionnaire in pregnancy and postpartum for screening, and it is also used as a continuous scale for symptom monitoring clinically and for research (HISCF, 2009; Howard et al., 2014). Scores on each EPDS item reflect the frequency of symptoms in the last two weeks and range from 0 to 3, with questions 3 and 5-10 reverse coded. Total scores range from 0 to 30. Higher scores indicate greater depressive symptomatology. As completing measures can be demanding, shortened versions with scores that perform comparably well with original full-length versions may help reduce the burden placed on respondents, as well as decrease the time it takes to administer the scale. However, shortening a scale is only advisable if it does not adversely affect measurement and screening accuracy properties of the scale.

Shortened forms of the full 10-item EPDS have been developed (Table 1) (Choi et al., 2012; Eberhard-Gran, Eskild, Samuelsen, & Tambs, 2007; Gollan et al., 2017; Martinez, Magana, Vohringer, Guajardo, & Rojas, 2020; Pallant, Miller, & Tennant, 2006; Venkatesh, Zlotnick, Triche, Ware, & Phipps, 2014). These include two two-item forms (Choi et al., 2012; Venkatesh et al., 2014), a five-item form (Eberhard-Gran et al., 2007), three- and seven-item subscales that measure symptoms of anxiety and depression separately (Gollan et al., 2017; Venkatesh et al., 2014), a three-item form (Martinez et al., 2020), and an eight-item form (Pallant et al., 2006). None of the development processes for these shortened forms used pre-specified criteria for performance to determine how many items to remove from the full 10-item EPDS. Furthermore, only three studies shortening the EPDS validated against major depression classification status (Eberhard-Gran et al., 2007; Martinez et al., 2020; Venkatesh et al., 2014), and these studies included only 63, 19, and 9 major depression cases. The extent to which the existing shortened forms retain the measurement and diagnostic properties of the full scale is unclear. Individual participant data meta-analysis (IPDMA), in which participant-level data from many studies are synthesized, allows for the development of a shortened form using data from a large number of participants.

Optimal test assembly (OTA) is a mixed-integer programming procedure that uses an estimated item response theory (IRT) model to select the subset of items that maximizes performance with respect to a given metric while satisfying pre-specified constraints (Linden, 2005). While more commonly used in the development of high-stakes educational tests (Kuhn & Kiefer, 2013), OTA is being increasingly used to develop shortened versions of patient-reported outcome measures (D. Harel et al., 2019; Ishihara et al., 2019; A. W. Levis et al., 2016). This

procedure was also shown to be replicable, reproducible, and to produce shortened forms of minimal length compared to alternative methods (D. Harel & Baron, 2019).

Aims of the Study

The objective of the present study was to apply optimal test assembly methods to a large database in order to develop a shortened version of the Edinburgh Postnatal Depression Scale. We (1) used confirmatory factor analysis to verify the unidimensionality of the underlying construct measured by the Edinburgh Postnatal Depression Scale; (2) applied optimal test assembly methods to obtain candidate forms of each possible length; and (3) selected the shortest possible form that showed similar performance to the full form in terms of pre-specified validity, reliability, and screening accuracy criteria, compared to the Edinburgh Postnatal Depression Scale.

MATERIALS AND METHODS

This study used a subset of data accrued for an IPDMA on the diagnostic accuracy of the EPDS for screening to detect major depression among pregnant and postpartum women. This IPDMA was registered in PROSPERO (CRD42015024785) and a protocol was published (B. D. Thombs et al., 2015). The protocol for the main IPDMA did not include methods for the present study. A protocol for the present study was uploaded to the Open Science Framework repository prior to initiating the study (https://osf.io/3cepr/).

Study Eligibility for the Main IPDMA

Datasets from articles in any language were eligible if they included women ≥ 18 years who were pregnant or had given birth in the previous year and both: (a) EPDS scores and (b) diagnostic classification for a current Major Depressive Episode (MDE) using Diagnostic and

Statistical Manual of Mental Disorders (DSM) or International Classification of Diseases (ICD) criteria based on a validated semi-structured or fully structured interview, administered within two weeks of each other. Participants recruited from psychiatric settings or setting where scales or interviews were administrated because of reported symptoms of depression were excluded, since screening is done to identify previously unrecognized cases (B. Thombs et al., 2011). Not all participants in a dataset needed to be eligible, if primary data allowed the selection of eligible participants.

Database Searches and Study Selection

A medical librarian searched Medline, Medline In-Process & Other Non-Indexed Citations and PsycINFO via OvidSP, and Web of Science Core Collections via ISI Web of Knowledge from inception to October 3, 2018, using a peer-reviewed (Sampson, McGowan, Lefebvre, Moher, & Grimshaw, 2008) search strategy (SupplementaryMethods1). We reviewed reference lists of relevant reviews and queried contributing authors about non-published studies. Search results were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA). After de-duplication, remaining citations were uploaded into DistillerSR (Evidence Partners, Ottawa, Canada) for processing review results.

Two investigators independently reviewed titles and abstracts. If either deemed a study potentially eligible, full-text review was done by two investigators, independently, with disagreements resolved by consensus, consulting a third investigator when necessary.

Data Contribution, Extraction, and Synthesis

Authors of eligible datasets were invited to contribute de-identified primary data, including EPDS item scores and major depression status. We emailed corresponding authors of

eligible primary studies at least three times, as necessary. If there was no response, we emailed co-authors and attempted phone contact.

Individual participant data were converted to a standard format and synthesized into a single dataset. We compared published participant characteristics and accuracy results with results from raw datasets and resolved any discrepancies in consultation with primary investigators.

For defining major depression, we considered MDD or MDE based on the DSM or ICD. If more than one was reported, we prioritized MDE over MDD. This is because screening would attempt to detect depressive episodes; further interview would determine if the episode is related to MDD, bipolar disorder, or persistent depressive disorder. We also prioritized DSM over ICD.

When datasets included statistical weights to reflect sampling procedures, we used the provided weights. For studies where sampling procedures merited weighting (e.g., all participants with positive screens and a random subset of participants with negative screens received a diagnostic interview), but the original study did not weight, we used inverse selection probabilities.

Data Eligibility for Present Study

For the present study, from the main IPDMA dataset, we only included primary studies that classified major depression based on the Structured Clinical Interview for DSM (SCID) (First, 2014). The SCID is a semi-structured diagnostic interview that was designed to be conducted by experienced diagnosticians. It requires clinical judgment and allows rephrasing questions and probes to follow up responses. Fully structured interviews, on the other hand, are fully scripted, with no allowance for deviation from the script. These interviews remove clinical judgement from the process, allowing lay interviewers, rather than clinicians, to perform the

assessment. Because of this, they may sacrifice validity. In recent analyses using three large IPDMA databases (B. Levis et al., 2018; B. Levis et al., 2019; Wu et al., 2020), it was found that compared to semi-structured interviews, fully structured interviews, which are designed for administration by lay interviewers, may identify more patients with low-level symptoms as depressed but fewer patients with high-level symptoms. Furthermore, a very brief version, the Mini International Neuropsychiatric Interview, identified far more participants as being depressed across the symptom spectrum (B. Levis et al., 2018; B. Levis et al., 2019; Wu et al., 2020). These results were consistent with the idea that semi-structured interviews most closely replicate clinical interviews done by trained professionals, whereas fully structured interviews are less rigorous reference standards. They are less resource-intensive options that can be administered by research staff without diagnostic skills but may misclassify major depression in substantial numbers of patients. Semi-structured interviews replicate diagnostic standards more closely than other types of interviews, and the SCID is by far the most commonly used semi-structured diagnostic interview for depression research [34-36]. In our main EPDS IPDMA database, 34 of 36 studies that used semi-structured interviews to classify major depression status used the SCID. Therefore, we only included SCID studies.

In addition, as EPDS item-level data was necessary for the proposed analyses, we only included studies in which EPDS item-level data (not just total scores) were available. For studies that collected data at multiple time points, we selected the time point with the most participants. If there was a tie, we selected the time point with the most major depression cases.

Statistical Analyses

All analyses were conducted using R version 3.6.0.

Verification of Unidimensionality of the EPDS

Robust weighted least squares estimation in R was used to fit a single-factor confirmatory factor analysis model of EPDS items (Muthen & Muthen, 1998). The model was first fit without allowing for any residual correlations among the items. If there was poor model fit, and if warranted by theoretical justification, modification indices were to be used to identify item pairs that would improve model fit by allowing their residuals to correlate (McDonald & Ho, 2002). Model fit was evaluated concurrently, using the χ^2statistic, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA) (Chen, Curran, Bollen, Kirby, & Paxton, 2008). Priority was given to CFI, TLI, and RMSEA, because the χ^2test may reject well-fitting models when sample size is large (Reise, Widaman, & Pugh, 1993). Model fit was considered to be adequate if CFI and TLI were ≥ 0.95 and RMSEA ≤ 0.08 (Hu & Bentler, 1999). The confirmatory factor analysis was fit using the **lavaan** package (Rosseel, 2012).

Item Response Theory Model and Optimal Test Assembly

A generalized partial credit model (GPCM) was fit to EPDS pooling data from all included studies (Muraki, 1992). The GPCM is an IRT model that relates a latent trait, representing severity of depressive symptomatology, to the distribution of observed item-level responses. The GPCM estimates two types of item-specific parameters: a discrimination parameter and threshold parameters. From these item-level parameter estimates, item information functions for each item were calculated from the GPCM, as well as a test information function (TIF), obtained by summing item information functions. Because the TIF is inversely related to the standard error of measurement of the latent trait, high amounts of information represent greater precision for measuring depressive symptomatology. The GPCM was fit using the **ltm** package (Rizopoulos, 2006).

Next, we used OTA - a mixed-integer programming technique - to systematically search for the short form that maximized the TIF, subject to the constraint of fixing the number of items included in each short form. By using the TIF as the objective function, the procedure optimizes the precision of the short form in estimating participants' level of depressive symptomatology (Linden, 2005; van ver Linden & Boekkooi-Timminga, 1989). The shape of the TIF was anchored at five points (Linden, 2005). Thus, for each short form of lengths 1 to 9 items, OTA selected items from the full set of EPDS items that maximized the test information. The OTA analysis was conducted using the **lpSolveAPI** package.

For each of the 9 candidate short forms and the full-length form, two scoring procedures were used to obtain estimates of each participant's level of depressive symptomatology. First, the summed scores across all items included in the short form were calculated. Second, factor scores were estimated for each participant. Although summed scores are typically relied upon for clinical use, the factor scores are considered to provide a better estimate of the latent trait due to well-known limitations of the summed score under the GPCM (Daphna Harel, 2014; Van der Ark, 2005).

Selection of Final Short Form

The elimination of items necessarily reduces information compared to a full-length form. Thus, to guarantee adequate performance, the selection of the final short form was based on the following five criteria: reliability, concurrent validity of summed scores, concurrent validity of factor scores, and non-inferior sensitivity and specificity.

Reliability of each candidate short form was assessed with Cronbach's alpha (Cronbach, 1951), since it is commonly used in research, despite limitations. The final selected form was required a priori to have a Cronbach's alpha coefficient ≥ 0.80. Concurrent validity of the

summed scores and factor scores was measured with the Pearson's correlation coefficient between candidate short form scores and the full-length EPDS. It was required a priori to be ≥ 0.90 (D. Harel & Baron, 2019).

Diagnostic accuracy of each candidate short form was assessed through a three-step process. First, pooled sensitivity and specificity of each candidate short form (compared to the SCID) for each of its possible cutoff summed score values were estimated with a bivariate random-effects model. Second, for each candidate short form, an optimal cutoff score was selected using Youden's J statistic (sensitivity + specificity -1) (B. Levis, Negeri, Sun, Benedetti, & Thombs, 2020; Youden, 1950). The bivariate random-effects model was fit using the **lme4** package (Bates, Machler, Bolker, & Walker, 2015).

Third, two non-inferiority tests were conducted for each of the 9 candidate forms to compare sensitivity and specificity, separately, to the full-length form. Non-inferiority tests assess whether the sensitivity or specificity of the short form is not lower than that of the full-length form, up to a pre-specified clinically significant tolerance of $\delta = 0.05$ (Counsell & Cribbie, 2015). To conduct the non-inferiority test, the sampling distribution of the test statistic was generated through the bootstrap method (Liu, Ma, Wu, & Tai, 2006). Bootstrapping resamples the original dataset with replacement to generate new, artificial, datasets (Efron & Tibshirani, 1994). For each non-inferiority test, 2000 bootstrap iterations were conducted, controlling in each for the number of respondents with and without major depression. For each bootstrap iteration, the bivariate random-effects model was fit to each of the 9 candidate short forms and the full-length form, and the sensitivities and specificities were computed based on their cutoff scores. To account for the multiple testing in the 18 total non-inferiority tests,

Benjamini-Hochberg adjusted p-values were used to determine the significance of the tests at the 0.05 significance level (Benjamini & Hochberg, 1995).

Funding and ethics

The study sponsors had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication. DH had full access to all data in the study and had final responsibility for the decision to submit for publication. As this study involved secondary analysis of de-identified previously collected data, the Research Ethics Committee of the Jewish General Hospital declared that this project did not require research ethics approval. However, for each included dataset, we confirmed that the original study received ethics approval and that all patients provided informed consent.

RESULTS

Search Results and Inclusion of Primary Data

Of 4,434 unique titles and abstracts identified from the database search, 4,056 were excluded after title and abstract review and 257 after full-text review, leaving 121 eligible articles with data from 81 unique participant samples, of which 56 (69%) contributed datasets (SupplementaryFigure1). Authors of included studies contributed data from two additional studies that were not retrieved by the search, for a total of 58 datasets. Of these, we excluded 24 studies that used a diagnostic interview other than the SCID and 12 more studies that did not have EPDS item scores available. In total, 5,157 participants (765 major depression cases) from 22 primary studies were included. These studies were conducted in 18 different countries, with 17 different languages. The mean age of the sample was 29.1 years. See Table 2 for descriptive sample statistics, and SupplementaryTable1 for characteristics of each included study.

Unidimensionality of the EPDS

A single factor model was fit to the EPDS-10 with residuals modeled as uncorrelated ($\chi^2[df = 65] = 663.1, p < 0.0001, TLI = 0.992, CFI = 0.988, RMSEA = 0.042$). As this model was deemed to be well fitting, no modification indices were used. Factor loadings for items were all high, with a median of 0.97 and a range of 0.88 to 1.15.

Item Response Theory Model and Optimal Test Assembly

The discrimination parameters for each item based on the GPCM are presented in Table 3. The information functions of each of the 10 items, as well as the total TIF are shown in Figure 1. The item with the greatest discrimination parameter was item 8, and thus has the most peaked information function in Figure 1. Other items with high values of the discrimination parameter and peaked information functions were items 1, 2 and 9. Table 4 shows the items that were included in each of the 9 candidate short forms from the OTA analysis. Item 8 was included in all candidate short forms, with items 3, 5, and 6 quickly dropped.

Selection of final short form

Cronbach's alpha values and concurrent validity correlations for the 9 candidate short forms are presented in Table 5. The results of the non-inferiority tests for both sensitivity and specificity are presented in Table 6.

The 5-item short form (EPDS-Dep-5) was the shortest form that fulfilled all criteria. The form included item 1 ("I have been able to laugh and see the funny side of things"), item 2 ("I have looked forward with enjoyment to things"), item 8 ("I have felt sad or miserable"), item 9 ("I have been so unhappy that I have been crying"), and item 10 ("The thought of harming myself has occurred to me"). The EPDS-Dep-5 maintained high reliability with a Cronbach's alpha of 0.82 (95% CI, 0.81, 0.83) compared to 0.88 (95% CI, 0.87, 0.88) for the full-length

449 form. Correlations of the summed and factor scores between the EPDS-Dep-5 and EPDS-10
450 were 0.91 (95% CI, 0.91, 0.92) and 0.95 (95% CI, 0.91, 0.97), respectively. Youden's J for the
451 full EPDS and EPDS-Dep-5, at their optimal cutoffs of 11 or greater and 4 or greater,
452 respectively, were both 0.68. Receiver operating curves for the full EPDS and EPDS-Dep-5 are
453 presented in SupplementaryFigure 2. The sensitivity and specificity of the EPDS-Dep-5 at its
454 optimal cutoff of 4 or greater were 0.83 (95% CI, 0.73, 0.89) and 0.86 (95% CI, 0.80, 0.90),
455 respectively. Both sensitivity and specificity were non-inferior to the sensitivity (0.80; 95% CI,
456 0.71, 0.86) and specificity (0.88; 95% CI, 0.83, 0.92) of the full-length form.

457 **DISCUSSION**

458 This study used OTA to shorten the EPDS to a 5-item shortened version (EPDS-Dep-5)
459 while maintaining comparable measurement properties and screening accuracy to detect major
460 depression among women in pregnancy and postpartum. The implication of this research is that
461 shortening this scale allows for shorter administration times and places lower burden on
462 respondents without significantly reducing the ability of the scale to measure depressive
463 symptomology.

464 The EPDS-Dep-5 maintained similar sensitivity and specificity to that of the full-length
465 form and resulted in a minimal loss of information. Furthermore, the shortened form maintained
466 reliability and validity that were comparable to the full-length form based on pre-specified
467 criteria. Cronbach's alpha of the EPDS-Dep-5 was within 0.06 of that for the full-length form,
468 and correlations of the summed score and factor scores of the EPDS- 5 and EPDS-10 were 0.91
469 and 0.95. Per pre-specified criteria, the sensitivity and specificity of the EPDS-Dep-5 (0.825 and
470 0.859, respectively) were non-inferior to those of the EPDS-10 (0.797 and 0.880, respectively).

The 5 items included in the EPDS-Dep-5 included items 1, 2, 8, 9 and 10 from the original EPDS. These items cover the two core symptoms of depression – low mood (items 8 and 9) and anhedonia (items 1 and 2), as well as self-harm (item 10). Of note, although they were included as potential items for the final shortened form, none of the 3 anxiety items (items 3 [blame], 4 [anxious], and 5 [scared]) were retained in the EPDS-Dep-5. Our short form selection procedure assessed screening accuracy for detecting depression, not anxiety, and short form development for that purpose would need to be done separately.

Most existing studies developing shortened EPDS forms compared the shortened forms to the full EPDS rather than comparing to diagnostic classification for depression. Only three studies validated their shortened forms against major depression classification based on DSM or ICD diagnostic criteria, but these studies included only 63, 19 and 9 major depression cases (Eberhard-Gran et al., 2007; Martinez et al., 2020; Venkatesh et al., 2014), limiting their ability to draw conclusions about the shortened scales' measurement properties. Table 1 presents the items included in each study's shortened form as well as the methods used to create that version. The development of the EPDS-Dep-5 in the present study used data that originated from an IPDMA thus (1) providing the largest total sample size (5,157 participants), as well as data from multiple settings and countries, (2) used by far the largest number of major depression cases (765 cases), (3) used a validated semi-structured diagnostic interview as the reference standard for major depression classification (the SCID), and (4) used screening accuracy as part of the development process, not solely as a tool for validation. It was also the only study that used objective, pre-specified criteria for empirical selection of items to include in the short form.

This study showed that an EPDS-Dep-5 cutoff ≥ 4 maximized combined sensitivity and specificity using Youden's J (Youden, 1950). However, clinicians and researchers may consider

use of a higher cutoff if their goal is to only capture patients with high depressive symptom levels or a lower cutoff if their goal is to avoid false negatives.

There are several limitations for this study that must be considered. First, for the collection of data for the full IPDMA, it was not possible to obtain primary data from 25 of the 81 eligible datasets. In addition, of the 34 studies using the SCID that provided data for the full IPDMA, 12 did not provide EPDS item scores and thus could not be included in the present study. Second, although we included data from 22 studies that fulfilled strict inclusion criteria, including the use of the rigorous semi-structured SCID interview, there was still substantial heterogeneity across studies in terms of country and language which both allows for the generalization of the results to larger and more diverse populations but also may not select the optimal shortened form for each individual context. Third, the present study did not conduct a risk of bias assessment, however the full IPDMA from which a subset of data was selected for this study did conduct a risk of bias assessment using QUADAS-2. No QUADAS-2 domain items were consistently associated with differences in sensitivity or specificity estimates. Furthermore, the OTA procedure, is a data-driven approach, and therefore the results of this study should be replicated or cross-validated. Lastly, future work may consider assessing whether the EPDS-Dep-5 is subject to issues of poor item fit or differential item functioning.

CONCLUSION

The study used the OTA method to develop a valid and reliable 5-item shortened form of the EPDS using pre-specified objective criteria to determine the length and items included in the EPDS-Dep-5. This method was implemented with a sample of 5,157 participants from 22 primary studies. The resulting 5-item shortened version maintained measurement properties and screening accuracy of the full-length form within pre-specified limits.

References

Bates, D., Machler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 57*(1), 289-300.

Bennett, H. A., Einarson, A., Taddio, A., Koren, G., & Einarson, T. R. (2004). Prevalence of depression during pregnancy: Systematic review. *Obstetrics and Gynecology, 103*(4), 698-709. doi:10.1097/01.AOG.0000116689.75396.5f

Canada, M. H. C. o. (2012). Changing directions, changing lives: The mental health strategy for Canada. In: Mental Health Commission of Canada Calgary, AB.

Chen, F. N., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*(4), 462-494. doi:10.1177/0049124108314720

Choi, S. K., Kim, J. J., Park, Y. G., Ko, H. S., Park, I. Y., & Shin, J. C. (2012). The Simplified Edinburgh Postnatal Depression Scale (EPDS) for Antenatal Depression: Is It a Valid Measure for Pre-Screening? *International Journal of Medical Sciences, 9*(1), 40-46. doi:DOI 10.7150/ijms.9.40

Cooper, P. J., Campbell, E. A., Day, A., Kennerley, H., & Bond, A. (1988). Non-Psychotic Psychiatric-Disorder after Childbirth - a Prospective-Study of Prevalence, Incidence, Course and Nature. *British Journal of Psychiatry, 152*, 799-806. doi:DOI 10.1192/bjp.152.6.799

Counsell, A., & Cribbie, R. A. (2015). Equivalence tests for comparing correlation and regression coefficients. *Br J Math Stat Psychol, 68*(2), 292-309. doi:10.1111/bmsp.12045

Cox, J. L., Murray, D., & Chapman, G. (1993). A Controlled-Study of the Onset, Duration and Prevalence of Postnatal Depression. *British Journal of Psychiatry, 163*, 27-31. doi:DOI 10.1192/bjp.163.1.27

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

Duhoux, A., Fournier, L., Gauvin, L., & Roberge, P. (2013). What is the association between quality of treatment for depression and patient outcomes? A cohort study of adults consulting in primary care. *Journal of Affective Disorders, 151*(1), 265-274. doi:10.1016/j.jad.2013.05.097

Duhoux, A., Fournier, L., Nguyen, C. T., Roberge, P., & Beveridge, R. (2009). Guideline concordance of treatment for depressive disorders in Canada. *Social Psychiatry and Psychiatric Epidemiology, 44*(5), 385-392. doi:10.1007/s00127-008-0444-8

Eberhard-Gran, M., Eskild, A., Samuelsen, S. O., & Tambs, K. (2007). A short matrix-version of the Edinburgh Depression Scale. *Acta Psychiatrica Scandinavica, 116*(3), 195-200. doi:10.1111/j.1600-0447.2006.00934.x

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*: CRC press.

First, M. B. (2014). Structured clinical interview for the DSM (SCID). *The encyclopedia of clinical psychology*, 1-6.

Gavin, N. I., Gaynes, B. N., Lohr, K. N., Meltzer-Brody, S., Gartlehner, G., & Swinson, T. (2005). Perinatal depression - A systematic review of prevalence and incidence. *Obstetrics and Gynecology, 106*(5), 1071-1083. doi:DOI 10.1097/01.AOG.0000183597.31630.db

Gaynes, B. N., Gavin, N., Meltzer-Brody, S., Lohr, K. N., Swinson, T., Gartlehner, G., . . . Miller, W. C. (2005). Perinatal Depression: Prevalence, Screening Accuracy, and Screening Outcomes: Evidence Report/Technology Assessment, Number 119.

Gollan, J. K., Wisniewski, S. R., Luther, J. F., Eng, H. F., Dills, J. L., Sit, D., . . . Wisner, K. L. (2017). Generating an efficient version of the Edinburgh Postnatal Depression Scale in an urban obstetrical population. *Journal of Affective Disorders, 208*, 615-620. doi:10.1016/j.jad.2016.10.013

Harel, D. (2014). *The effect of model misspecification for polytomous logistic adjacent category item response theory models.* McGill University Libraries,

Harel, D., & Baron, M. (2019). Methods for shortening patient-reported outcome measures. *Stat Methods Med Res, 28*(10-11), 2992-3011. doi:10.1177/0962280218795187

Harel, D., Mills, S. D., Kwakkenbos, L., Carrier, M. E., Nielsen, K., Portales, A., . . . Investigators, S. (2019). Shortening patient-reported outcome measures through optimal test assembly: application to the Social Appearance Anxiety Scale in the Scleroderma Patient-centered Intervention Network Cohort. *BMJ Open, 9*(2), e024010. doi:10.1136/bmjopen-2018-024010

HISCF, A. (2009). Alberta Postpartum Depression-Data Set.

Howard, L. M., Molyneaux, E., Dennis, C. L., Rochat, T., Stein, A., & Milgrom, J. (2014). Perinatal mental health 1 Non-psychotic mental disorders in the perinatal period. *Lancet, 384*(9956), 1775-1788. doi:Doi 10.1016/S0140-6736(14)61276-9

Hu, L. T., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling-a Multidisciplinary Journal, 6*(1), 1-55. doi:10.1080/10705519909540118

Ishihara, M., Harel, D., Levis, B., Levis, A. W., Riehm, K. E., Saadat, N., . . . Thombs, B. D. (2019). Shortening self-report mental health symptom measures through optimal test assembly methods: Development and validation of the Patient Health Questionnaire-Depression-4. *Depress Anxiety, 36*(1), 82-92. doi:10.1002/da.22841

Kessler, R. C. (2003). Epidemiology of women and depression. *J Affect Disord, 74*(1), 5-13. doi:10.1016/s0165-0327(02)00426-3

Kuhn, J. T., & Kiefer, T. (2013). Optimal Test Assembly in Practice The Design of the Austrian Educational Standards Assessment in Mathematics. *Zeitschrift Fur Psychologie-Journal of Psychology, 221*(3), 190-200. doi:10.1027/2151-2604/a000146

Levis, A. W., Harel, D., Kwakkenbos, L., Carrier, M. E., Mouthon, L., Poiraudeau, S., . . . the Scleroderma Patient-Centered Intervention Network, I. (2016). Using Optimal Test Assembly Methods for Shortening Patient-Reported Outcome Measures: Development and Validation of the Cochin Hand Function Scale-6: A Scleroderma Patient-Centered Intervention Network Cohort Study. *Arthritis Care Res (Hoboken), 68*(11), 1704-1713. doi:10.1002/acr.22893

Levis, B., Benedetti, A., Riehm, K. E., Saadat, N., Levis, A. W., Azar, M., . . . Thombs, B. D. (2018). Probability of major depression diagnostic classification using semi-structured versus

603 fully structured diagnostic interviews. *British Journal of Psychiatry, 212*(6), 377-385.
604 doi:10.1192/bjp.2018.54
605 Levis, B., McMillan, D., Sun, Y., He, C., Rice, D. B., Krishnan, A., . . . Thombs, B. D. (2019).
606 Comparison of major depression diagnostic classification probability using the SCID,
607 CIDI, and MINI diagnostic interviews among women in pregnancy or postpartum: An
608 individual participant data meta-analysis. *International Journal of Methods in Psychiatric*
609 *Research*. doi:ARTN e1803
610 10.1002/mpr.1803
611 Levis, B., Negeri, Z., Sun, Y., Benedetti, A., & Thombs, B. D. (2020). Accuracy of the Edinburgh
612 Postnatal Depression Scale (EPDS) for screening to detect major depression: systematic
613 review and meta-analysis of individual participant data. *BMJ, 371*.
614 Linden, W. J. v. d. (2005). *Linear models of optimal test design*. New York, NY: Springer.
615 Liu, J. P., Ma, M. C., Wu, C. Y., & Tai, J. Y. (2006). Tests of equivalence and non-inferiority for
616 diagnostic accuracy based on the paired areas under ROC curves. *Statistics in Medicine,*
617 *25*(7), 1219-1238. doi:10.1002/sim.2358
618 Martinez, P., Magana, I., Vohringer, P. A., Guajardo, V., & Rojas, G. (2020). Development and
619 validation of a three-item version of the Edinburgh Postnatal Depression Scale. *J Clin*
620 *Psychol, 76*(12), 2198-2211. doi:10.1002/jclp.23041
621 McDonald, R. P., & Ho, M. H. (2002). Principles and practice in reporting structural equation
622 analyses. *Psychol Methods, 7*(1), 64-82. doi:10.1037/1082-989x.7.1.64
623 Muraki, E. (1992). A Generalized Partial Credit Model - Application of an Em Algorithm. *Applied*
624 *Psychological Measurement, 16*(2), 159-176. doi:Doi 10.1177/014662169201600206
625 Muthen, L., & Muthen, B. (1998). Mplus User's Guide, Statistical Analysis With Latent Variables.
626 *Confirmatory Factor Analysis and Structural Equation Modeling, 2012*(55), 111.
627 OHara, M. W., & Swain, A. M. (1996). Rates and risk of postpartum depression - A meta-
628 analysis. *International Review of Psychiatry, 8*(1), 37-54. doi:Doi
629 10.3109/09540269609037816
630 Ohara, M. W., Zekoski, E. M., Philipps, L. H., & Wright, E. J. (1990). Controlled Prospective-Study
631 of Postpartum Mood Disorders - Comparison of Childbearing and Nonchildbearing
632 Women. *Journal of Abnormal Psychology, 99*(1), 3-15. doi:Doi 10.1037/0021-843x.99.1.3
633 Pallant, J. F., Miller, R. L., & Tennant, A. (2006). Evaluation of the Edinburgh Post Natal
634 Depression Scale using Rasch analysis. *Bmc Psychiatry, 6*. doi:Artn 28
635 10.1186/1471-244x-6-28
636 Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item
637 response theory: two approaches for exploring measurement invariance. *Psychol Bull,*
638 *114*(3), 552-566. doi:10.1037/0033-2909.114.3.552
639 Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory
640 analyses. *Journal of Statistical Software, 17*(5).
641 Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical*
642 *Software, 48*(2), 1-36.
643 Sampson, M., McGowan, J., Lefebvre, C., Moher, D., & Grimshaw, J. (2008). PRESS: peer review
644 of electronic search strategies. *Ottawa: Canadian Agency for Drugs and Technologies in*
645 *Health*.

Stewart, D. E. (2011). Depression during Pregnancy. *New England Journal of Medicine, 365*(17), 1605-1611. doi:DOI 10.1056/NEJMcp1102730

Thombs, B., Arthurs, E., El-Baalbaki, G., Meijer, A., Ziegelstein, R., & Steele, R. (2011). Risk of Bias from the Inclusion of Already Diagnosed or Treated Patients in Diagnostic Accuracy Studies of Depression Screening Tools. *American Journal of Epidemiology, 173*, S324-S324.

Thombs, B. D., Benedetti, A., Kloda, L. A., Levis, B., Riehm, K. E., Azar, M., . . . Vigod, S. (2015). Diagnostic accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for detecting major depression in pregnant and postnatal women: protocol for a systematic review and individual patient data meta-analyses. *BMJ Open, 5*(10), e009742. doi:10.1136/bmjopen-2015-009742

Van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70*(2), 283-304. doi:10.1007/s11336-000-0862-3

van ver Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for IRT-based test design with practical constraints. *Psychometrika, 54*(2), 237-247.

Venkatesh, K. K., Zlotnick, C., Triche, E. W., Ware, C., & Phipps, M. G. (2014). Accuracy of Brief Screening Tools for Identifying Postpartum Depression Among Adolescent Mothers. *Pediatrics, 133*(1), E45-E53. doi:10.1542/peds.2013-1628

Vesga-Lopez, O., Blanco, C., Keyes, K., Olfson, M., Grant, B. F., & Hasin, D. S. (2008). Psychiatric disorders in pregnant and postpartum women in the United States. *Archives of General Psychiatry, 65*(7), 805-815. doi:DOI 10.1001/archpsyc.65.7.805

Whitley, R., & Kirmayer, L. J. (2008). Perceived stigmatisation of young mothers: An exploratory study of psychological and social experience. *Social Science & Medicine, 66*(2), 339-348. doi:10.1016/j.socscimed.2007.09.014

Wu, Y., Levis, B., Sun, Y., Krishnan, A., He, C., Riehm, K. E., . . . Thombs, B. D. (2020). Probability of major depression diagnostic classification based on the SCID, CIDI and MINI diagnostic interviews controlling for Hospital Anxiety and Depression Scale - Depression subscale scores: An individual participant data meta-analysis of 73 primary studies. *J Psychosom Res, 129*, 109892. doi:10.1016/j.jpsychores.2019.109892

Youden, W. J. (1950). Index for Rating Diagnostic Tests. *Cancer, 3*(1), 32-35. doi:Doi 10.1002/1097-0142(1950)3:1<32::Aid-Cncr2820030106>3.0.Co;2-3

Zelkowitz, P., & Milet, T. H. (1996). Postpartum psychiatric disorders: Their relationship to psychological adjustment and marital satisfaction in the spouses. *Journal of Abnormal Psychology, 105*(2), 281-285. doi:Doi 10.1037/0021-843x.105.2.281

Zelkowitz, P., & Milet, T. H. (2001). The course of postpartum psychiatric disorders in women and their partners. *Journal of Nervous and Mental Disease, 189*(9), 575-582. doi:Doi 10.1097/00005053-200109000-00002