Developing and validating a perinatal depression screening tool in Kenya blending Western criteria with local idioms: A mixed methods study

Eric P. Green
Duke University

Hawa Tuli
Duke University

Edith Kwobah
Moi Teaching and Referral Hospital

Menya D.
Moi University

Irene Chesire
Moi University

Christina Schmidt
Duke University

Abstract

Background: Routine screening for perinatal depression is not common in most primary health care settings. The U.S. Preventive Services Task Force only recently updated their recommendation on depression screening to specifically recommend screening during the pre- and postpartum periods. While practitioners in high-income countries can respond to this new recommendation by implementing one of several existing depression screening tools developed in Western contexts, such as the Edinburgh Postnatal Depression Scale (EPDS) or the Patient Health Questionnaire-9 (PHQ-9), these tools lack strong evidence of cross-cultural equivalence, validity for case finding, and precision in measuring response to treatment in developing countries. Thus, there is a critical need to develop and validate new screening tools for perinatal depression that can be used by lay health workers, primary health care personnel, and patients.

Methods: Working in rural Kenya, we used free listing, card sorting, and item analysis methods to develop a locally-relevant screening tool that blended Western psychiatric concepts with local idioms of distress. We conducted a validation study with a random sample of 193 pregnant women and new mothers to test the diagnostic accuracy of this scale along with the EPDS and PHQ-9.

Results: The sensitivity/specificity of the EPDS and PHQ-9 was estimated to be 0.70/0.72 and 0.70/0.73, respectively. This compared to sensitivity/specificity of 0.90/0.90 for a new 9-item locally-developed tool called the Perinatal Depression Screening (PDEPS). Across these three tools, internal consistency reliability ranged from 0.77 to 0.81 and test-retest reliability ranged from 0.57 to 0.67. The prevalence of depression ranges from 5.2% to 6.2% depending on the clinical reference standard.

Conclusion: The EPDS and PHQ-9 are valid and reliable screening tools for perinatal depression in rural Western Kenya, the PDEPS may be a more useful alternative. At less than 10%, the prevalence of depression in this region appears to be lower than other

published estimates for African and other low-income countries.

Developing and validating a perinatal depression screening tool in Kenya blending Western criteria with local idioms: A mixed methods study

Introduction

Depression is a leading cause of disability worldwide, yet access to timely assessment and treatment is very limited in many low-income settings, especially in rural communities. Depression affects men and women, young and old, but women who experience depression during pregnancy or in the year after childbirth are a particularly underserved population. The prevalence of perinatal depression among women living in poor countries ranges widely, possibly exceeding 30 percent in rural settings (Villegas, McKay, Dennis, & Ross, 2011).

Depression among pregnant women and new mothers has been linked to increased maternal morbidity and mortality (Khalifeh, Hunt, Appleby, & Howard, 2016; Oates, 2003), poor infant health (Field et al., 2004; Gelaye, Rondon, Araya, & Williams, 2016; Grigoriadis et al., 2013; Rahman et al., 2016; Surkan, Patel, & Rahman, 2016), and poor early childhood outcomes—such as developmental, cognitive, and emotional delays (Beck, 1998; Gentile, 2017; Junge et al., 2017)—making it a significant public health concern. Few public health systems currently have the resources to treat perinatal depression, but recent work has shown that cognitive behavioral interventions delivered by lay health workers are efficacious (Joshi et al., 2014; Rahman, Malik, Sikander, Roberts, & Creed, 2008). Before such treatments can be delivered at scale, however, it is essential to overcome many barriers, including barriers to screening for depression.

Routine screening for perinatal depression is not common in most primary health care settings. The U.S. Preventive Services Task Force only recently updated their recommendation on depression screening to specifically recommend screening during the pre- and postpartum periods (Albert L. Siu & the US Preventive Services Task Force, 2016). While practitioners in high-income countries can respond to this new recommendation by implementing one of several existing depression screening tools developed in Western contexts, such as the Edinburgh Postnatal Depression Scale (EPDS)

or the Patient Health Questionnaire-9 (PHQ-9), these tools lack strong evidence of cross-cultural equivalence, validity for case finding, and precision in measuring response to treatment in developing countries (Sweetland, Belkin, & Verdeli, 2014; Tsai et al., 2013). Thus, there is a critical need to develop and validate new screening tools for perinatal depression that can be used by lay health workers, primary health care personnel, and patients. Our study contributes to this effort by attempting to validate the EPDS and PHQ-9 in rural Kenya, while at the same time developing and validating a new instrument that blends items from existing screening tools with local idioms of distress (Kohrt et al., 2011).

Methods

Setting and Participants

We conducted this prospective study in Bungoma, Kenya. This rural county is situated in what used to be known as Western Province. When the 2010 Constitution of Kenya was enacted in 2013, 47 counties in a new devolved system of government replaced the existing 8 provinces. Bungoma is one of the largest counties in this new system. It is home to more than 1.6 million residents, nearly half of whom live in poverty (Wiesmann, Kiteme, & Mwangi, 2014).

We recruited participants for two main study activities: (i) eight focus group discussions to develop a locally-anchored set of screening items and (ii) individual assessments to narrow the set of items and validate the new measure and two existing screening tools. A purposive sample of 12 women were invited to participate in the focus group discussions; women were eligible to participate if they were at least 18 years old and receiving maternity services from a particular primary health clinic (public dispensary) in Bungoma East subcounty. All 38 community health volunteers (CHVs) serving the clinic's catchment area were invited to participate in separate discussion sessions.

For the validation study, we drew a random sample of 210 pregnant women and new

mothers (from a sampling frame of 276) from the 27 villages wholly or partially located within a 2-kilometer radius from the clinic. All women had to be at least 18 years of age. Pregnant women in their second or third trimesters were eligible, as were new mothers 1 to 6 months postpartum. Women who miscarried or experienced a stillbirth or infant death linked to their most recent pregnancy were excluded from the study. 193 women completed questionnaires and semi-structured clinical interviews.

Measures

Screening Survey (Index Tests). We identified 17 measures commonly used to assess perinatal depression (see Table A1 in the Online Appendix), created a database of 365 items, assigned every screening item a short cover term (e.g., crying, unhappy, heart racing), and reviewed each cover term for exact and approximate duplicates. Out of the initial 365 screening items, we identified 171 unique cover terms and wrote an index card (with English and Kiswahili writing) for each term in preparation for our focus group discussions. Through these discussions we created a 60-item survey that included several index tests: the Edinburgh Postnatal Depression Scale, the Patient Health Questionnaire-9, items from other existing screening tools, and new items generated by the focus groups. In addition to the screening items, the survey also included demographic questions from the Phase 6 and Phase 7 Demographic and Health Household and Woman's questionnaires (DHS Program, n.d.-a).

Edinburgh Postnatal Depression Scale. The most commonly used screening instrument for perinatal depression is the 10-item self-report Edinburgh Postnatal Depression Scale (EPDS; Cox, Holden, & Sagovsky, 1987). The first validation study was conducted with 84 postnatal women in the United Kingdom and reported sensitivity of 86 percent, specificity of 78 percent, and a positive predictive value of 73 percent. A systematic review of 37 EPSD validation studies conducted between 1987 and 2008, however, revealed great heterogeneity in diagnostic sensitivity and specificity between

studies for all cutoff points (Gibson, McKenzie-McHarg, Shakespeare, Price, & Gray, 2009).

Tsai et al. 2013 recently extended this evaluation of the EPDS with a new systematic review of 25 studies that screened for perinatal depression in Africa; 16 of the 25 studies included in this review used the EPDS. The authors noted that the median estimated coefficient alpha of the EPDS was 0.84, and they calculated a pooled sensitivity and specificity of 0.94 and 0.77 (cut-off ≥ 9) from 14 studies that assessed criterion validity. None of the included studies was conducted in Kenya.

Patient Health Questionnaire-9. Another brief depression screening that is often used to assess perinatal depression is the Patient Health Questionnaire-9 (PHQ-9). At least two studies have investigated Kiswahili translations of the PHQ-9 in Kenya. Omoro et al. 2006 demonstrated an association between PHQ-9 scores, TNM stage (Classification of Malignant Tumors), and scores on a cancer-specific quality of life scale. Monahan et al. 2009 found a correlation between scale scores and patient responses to the question, "In general how would you rate your overall health right now?" Neither study assessed validity by comparing results to a gold-standard, such as a clinical structured interview.

Criterion Reference: Structured Clinical Interview for DSM-5. We used the Structured Clinical Interview for DSM-5, Research Version to diagnose cases of depression (SCID-5-RV; First, Williams, Karg, & Spitzer, 2015). The SCID-5-RV is designed to be customized, and we opted to administer the non-patient overview, Module A on mood episodes with specifiers, Module BC for psychotic screening, and Module D for the differential diagnosis of mood disorders—all translated into Kiswahili prior to use. Table A2 in the Online Appendix details the modifications we made to each module.

The target condition was Major Depressive Episode (MDE). To meet criteria for a current MDE according to the DSM-5 (American Psychiatric Association, 2013), a woman had to experience at least 5 of 9 symptoms—including depressed mood (A1) or diminished interest or pleasure (A2)—during the same 2-week period within the past 1 month (Criterion A) and report that these symptoms caused clinically significant distress or

impairment in functioning (Criterion B). Four Kenyan counselors (2 Bachelor's-level, 2 Master's-level) investigated all cases in which a general medical condition, substance abuse, or medication could be the etiological factor (Criterion C). Counselors also used Module B/C to determine if psychotic symptoms were primarily accounted for by a DSM-5 Psychotic Disorder (Criterion D). Counselors did not assess Criterion E of Module D (i.e., rule out manic or hypomanic episode); therefore, we could only diagnose MDE not Major Depressive Disorder.

If a woman's symptoms suggested a depressive disorder but the woman did not meet Criterion A for MDE, the counselor assessed Criteria B-D to possibly diagnose "Other specified depressive disorder".

Alternate Criterion Reference: Local Diagnosis. In addition to using the SCID-5-RV to diagnose depression as defined by the DSM-5, we also asked counselors to use their clinical judgment and asked women to self-report on their well-being.

Clinical judgment of diagnosis and functioning. Counselors responded to the following prompt to record a 'local' diagnosis that was not tied to the DSM criteria: "In your clinical judgment, do you think that this woman is 'depressed'?" Counselors also rated each woman's social and occupational functioning using the SOFAS rating scale included in the SCID-RV-5. SOFAS ratings can range from 0 to 100, with 100 representing "superior functioning".

Client rating of functioning. Counselors asked women to rate their own well-being by indicating on which step of an imaginary 10-step ladder they stood. Women were told that people who were really struggling and not doing well stood on Step 1, and that people who were doing very well stood on Step 10.

Local reference standard: Concordance between counselor and client. If there was concordance between the counselor's local diagnosis of depression and the woman's report that she stood on steps 1 through 5 (not well dimension), we classified the woman as a 'local' case of depression (Bolton, 2001a).

Procedures

Figure 1 displays the overall study design and the participant flow for the diagnostic accuracy study.

Qualitative Study to Develop Potential Screening Items. We developed the set of screening items used in the validation study through a process of free listing and card sorting, expert review, and local adaptation. Our objective of this phase of research was to establish the content equivalence, content validity, semantic equivalence, and technical equivalence of the screening items (Kohrt et al., 2011).

Free listing and card sorting activity (content equivalence). Over the course of 7 days in June 2015, we conducted free listing and card sorting exercises with 2 groups of pregnant women and new mothers and 6 groups of CHVs. Free listing is an ethnographic research method that results in a list of responses to a single inquiry (Bolton, 2001b). After a brief discussion about the general maternal health issues in the community, we asked each group to list as many characteristics as they could think of in response to a prompt to describe what depression ("sadness" in Kiswahili) looks like in pregnant women and new mothers. A member of the research team who was fluent in both English and Kiswahili facilitated each discussion and probed for more details throughout the listing procedure. A second researcher took notes and created an index card for each characteristic mentioned by the group.

While the listing exercise was ongoing, a research assistant attempted to match the index cards generated by the group to our set of 171 cover terms (in the client focus groups we only attempted to match against the 54 most common cover terms). After recording direct matches, all non-matching cover term cards were spread out on the table for the group to review. We asked participants to sort the remaining Western psychiatric cards into four piles according to whether the characteristic is something that they observe in the setting: (i) "yes"; (ii) "no"; (iii) "maybe"; or (iv) "no opinion". We also tracked how frequently groups generated local constructs (idioms of distress) without matches in the set

of Western psychiatric cover terms (Kohrt et al., 2011). Low literacy levels were not an obstacle as some women in every group were literate.

Expert review (content validity). After the completion of the group discussions, a purposive sample of 11 mental health professionals at Moi University Teaching and Referral Hospital was invited by a Kenyan Co-Investigator to review the free listing and card sorting results. Participants included a psychiatrist, two clinical psychologists, and mental health ward nurses. The panel began the session with a discussion of the symptoms identified in the client and CHV free listing exercises, including cultural perceptions of depression and mental health. The expert panel also completed their own free listing activity, similar to the one performed in the client and CHV groups. The researchers attempted to match index cards generated by the panel to the 50 most highly endorsed cover terms from the previous focus groups, and then asked the experts to sort the remaining non-matching cover terms into the same four piles.

Item shortlisting and adaptation (semantic and technical equivalence). We created a 60-item screening that we administered to a random sample of pregnant women and new mothers in the validity study. In addition to the 10 EPDS items and the 9 PHQ-9 items, we identified 28 of the most endorsed cover terms and 5 of the most frequently mentioned local constructs. We developed each of these 33 terms into screening items that matched the format and response options of the PHQ-9. For instance, the cover term "temper" became "easily losing your temper" and followed the format of the PHQ-9 prompt: "For the past two weeks, how many times have you been bothered by the following problems?"

In addition to including the original 10 EPDS items, we also included revised EPDS items that matched the format and response options of the PHQ-9. First, we rephrased the new set of EPDS items so that women could indicate how often they have been bothered by the particular problem in the past two weeks, rather than the past one week (see Table A3 in the Online Appendix). Second, we reworded the two 'positive' EPDS items to

express problems. For example, EPDS item 2 is "I have looked forward with enjoyment to things." In addition to asking women to respond to this original EPDS item, we created a new version phrased as a problem, "Not looking forward to things".

One native speaker translated these 60 items from English to Kiswahili, and a second native speaker conducted the blind back-translation into English. A Kiswahili speaker on the research team resolved discrepancies and conducted cognitive interviewing on the translated items with a convenience sample of 14 pregnant women and new mothers attending the clinic (age M=26.0 years, SD=4.0 years). The team then adapted the items as necessary to enhance understanding and reviewed the final set with a Kenyan psychiatrist (Co-Investigator).

Quantitative Study to Validate Measures. Following the qualitative study, we recruited a probability sample of pregnant women and new mothers to complete three study activities: (i) a 60-item screening consisting of the EPDS, PHQ-9, and our new items; (ii) a re-test within 7 days; and (iii) a semi-structured clinical interview. Our objective of this phase of research was to select the final screening items and to establish the construct validity and diagnostic validity of these items (Kohrt et al., 2011).

Sampling. According to the 2014 Kenya DHS, 97.6 percent of women in Bungoma County who gave birth in the previous five years received antenatal care from a skilled provider (Kenya National Bureau of Statistics et al., 2015). So in theory, it should have been sufficient to use the antenatal register as a sampling frame for the population of pregnant women and new mothers living in the catchment area of the clinic. In practice, however, clinic registers can be incomplete and inaccurate. Therefore, working in collaboration with our clinic partner, we created a list of potentially eligible participants based on a review of the antenatal (ANC), postnatal (PNC), and delivery registers before conducting two verification exercises. A detailed participant flow is displayed in Figure 1.

We identified 711 study participants who were potentially eligible for a July 2015 study launch: (i) 308 women in the ANC register who were in their second or third

trimester, (ii) 245 women in the ANC register who, based on the recorded expected delivery date, should have given birth in the past 1 to 6 months; (iii) 58 women in the PNC register who attended well-baby checkups for babies born in the past 1 to 6 months; and (iv) 100 women in the labor and delivery register who delivered a baby at the facility in the past 1 to 6 months. 437 names remained after de-duplicating the list.

Next, CHVs for each of the 27 villages within a 2-kilometer radius of the clinic reviewed village-specific lists to verify accuracy and completeness. CHVs added a total of 134 names and removed 284 names of women who were ineligible, not living in the village, or duplicated in the list. The survey team then conducted a second verification exercise on the reduced list of 287 names and further trimmed a net of 11 names, resulting in a sampling frame of 276 women. We drew a simple random sample of 210 women from this sampling frame (see Appendix B for a description of the rationale for this sample size).

Screening and retest. Three Bachelor's-level Kenyan enumerators with a background in social work or mental health completed one day of training to administer the screening surveys using encrypted Android tablets running ODK Collect (version 1.4.5). Enumerators read the questions aloud and entered the participant's response. Completed data forms were sent daily to a secure research server for processing.

The same enumerator returned within 7 days to complete a re-test survey. Participants were randomized to complete the re-test the same way as before (i.e., the enumerator reading questions on the tablet and entering responses) or via a mobile phone using an automated interactive voice response system. For women who were assigned to the phone re-test, enumerators used a basic mobile phone to call the research number, entered the participant's study information, and then handed the phone to the participant. Women listened privately to automated prompts recorded in Kiswahili and pressed numbers on the phone keypad to respond.

Semi-structured clinical interview. Women also participated in a semi-structured clinical interview within 3 days of the index test to determine casesness.

The order of the survey and interview was randomized at the village level so that roughly half of participants completed the survey before the clinical interview, while the other half completed the interview before the survey. Interviewers and enumerators were blinded to the results and had limited contact with each other during the data collection period. Two Bachelor's-level counselors and two Master's-level counselors were trained by a clinical psychologist (Principal Investigator) and a Kenyan psychiatrist (Co-Investigator) to complete the SCID-5-RV. The team reviewed each section of the interview in detail and then the trainees took turns interviewing a research team member—a native Kiswahili speaker—who played the role of a target woman and followed scripted response sets designed to expose the clinicians to different interview scenarios. In these joint interviews, the non-interviewing trainee also recorded notes and ratings. The pair of clinicians reached 100 percent agreement on the diagnosis across 4 joint interviews.

The Master's-level counselors supervised the research effort and reviewed each completed SCID form and associated clinical notes. During the first week of interviews, the team met in person every evening to review case notes and ratings. Interviewers were trained to report all cases of current suicidal ideation, intent, or attempts, and severe cases of suspected MDE to supervisors for immediate review. With a woman's permission, the interviewer and supervisor made a referral to a counselor in the nearest town or, if necessary, the psychiatric nurse at the district hospital. The research team provided funds to ensure prompt transport and initial care.

Analysis

Qualitative Study to Develop Potential Screening Items. We constructed an overall endorsement score for each cover term by averaging the values assigned to each term during the focus group discussions with health workers. If a cover term matched a card generated by a group during the free listing process, the term was assigned a value of '4'. In each group, non-matching cover terms were presented for discussion. Non-matching

terms were assigned the following values: '3' if the group endorsed the characteristic, '2' if the group said that the term could be a possible fit, '0' if the group had no opinion, and '-1' if the group rejected the term. We ranked the cover terms by endorsement score and ranked local items without cover term matches by the frequency of mention by the groups. The expert panel reviewed the results with the research team and recommended adding and dropping items from our screening.

Quantitative Study to Validate Measures. Our empirical strategy consisted of two phases. First we examined which items did the best job discriminating between cases and non-cases and then assessed which combination of items optimized scale reliability and classification. Second, we assessed the diagnostic validity, construct validity, and reliability of the new scale and two existing scales.

Item analysis. We adopted an approach to item analysis used in the development of the General Health Questionnaire (Goldberg, 1972). All items had a 4-point response scale ranging from 0 to 3; higher numbers represented more endorsement of the symptom.[1] For each item, we calculated the proportion of participants who endorsed the item with a value of 2 or 3. We then subtracted the proportion of endorsement among non-cases from the proportion of endorsement among cases, resulting in a gradient score.

Items that do a better job discriminating between cases and non-cases have higher gradient scores; negative gradient scores indicate that a higher proportion of non-cases compared to cases endorsed the item. Therefore, we eliminated from further consideration items with a gradient score less than 0.05. We also eliminated items that were endorsed by more than 25 percent of non-cases since non-case endorsement suggests that the item measures a broader, or different construct than depression. The item analysis resulted in a subset of 20 of the original 60 items that we could further evaluate.

[1]The original 10 EPDS items had a different set of anchors than the other 50 screening items, but the values assigned to each EPDS response option still ranged from 0 to 3, with a value of 3 representing greater problems.

Item selection. Using the DSM-5 definition of caseness, we evaluated the internal consistency reliability (Cronbach's alpha) and classification accuracy of all 616,645 possible combinations of these 20 items in scales ranging in size from 2 to 10 items. We used the 'OptimalCutpoints' package (version 1.1-3; López-Ratón, Rodríguez-Álvarez, Suárez, & Sampedro, 2014) for R (version 3.2.3; R Core Team, 2015) to select the optimal cutpoint for each combination of items, giving equal weight to sensitivity and specificity in the identification of cases. We used these results to select the subset of items that would make up the new perinatal depression scale.

Diagnostic validity. We report the sensitivity, specificity, accuracy, positive and negative likelihood ratios, and area under the curve for the new perinatal depression scale in addition to the original EPDS and PHQ-9 scales. Confidence intervals for sensitivity and specificity measures are Rubin and Schenker's logit confidence intervals (see López-Ratón et al. 2014). One incomplete observation was dropped.

Construct validity. To assess convergent validity we calculated the correlation between each measure of depression severity and two measures of functioning: counselor SOFAS rating and client ladder rating. To assess discriminant validity we compared the mean score on each measure of depression severity by caseness.

Reliability. For the EPDS, PHQ-9, and the new scale, we estimated internal consistency reliability (Cronbach's alpha) for the first test. We also calculated test-retest reliability among the subset of women who completed a tablet re-retest survey within 7 days of the original tablet survey, and among the subset of women who completed a phone re-test survey within the same time period. We tested the null hypothesis of no difference between correlation coefficients by mode of re-test (enumerator administered survey vs automated phone survey) with the 'cocor' (Diedenhofen & Musch, 2015) package that implements Zou's (2007) method of calculating confidence intervals on the difference.

Demographics. We ran cross-tabulations of key demographic variables by maternal status and caseness. In constructing the wealth index from the DHS survey

questions we administered, we followed DHS guidance and used reference values from the Kenya 2008-09 DHS (DHS Program, n.d.-b).

Research Ethics

The study protocol was reviewed and approved by the Duke University Institutional Review Board and the Institutional Research and Ethics Committee in Kenya. All team members completed training in research ethics, and all study participants provided written informed consent. A completed STARD checklist is provided in Appendix C.

Results

Qualitative Study to Develop Potential Screening Items

Participant Characteristics. We conducted free listing and card sorting exercises with 2 groups of pregnant women and new mothers (n=12) and 6 groups of CHVs (n=38). On average, groups had 6.2 participants (SD=0.9). The average age of the female clients and CHVs was 28.2 (SD=3.4) and 41.4 years (SD=7.8), respectively. 84.2 percent of CHVs were female, and 55.3 percent finished secondary school. This compared to 33.3 percent of the female clients.

Item Shortlisting. The 6 groups of CHVs generated a total of 153 cards (25.5 cards per group; SD=3.2).[2] Overall, 58.8 percent of the these cards matched 1 of the 171 Western psychiatric cover terms, supporting the idea that many aspects of depression are universal. We calculated average endorsement scores for each Western cover term and found that the results from CHVs and female clients were highly correlated, $r(52) = 0.70$, $p < 0.001$. Since CHVs were asked to discuss the full set of 171 Western psychiatric cover terms, we used their average ratings to select a subset of the most endorsed terms to present to the expert panel, along with several of the most frequently mentioned local idioms of distress not represented in the set of Western terms. The expert panel reviewed

[2] As expected, the 2 groups of women generated fewer terms on average compared to the health worker groups (M=14.5; SD=2.1) and 124 fewer terms overall.

and approved of the qualitative results, but added one item for testing: “feeling like you just want to go back to your maternal home”, a reference to the tradition that women move to the husband’s home area upon marriage. In total, we selected 60 items for inclusion in the validation survey.

Quantitative Study to Validate Measures

Participant Characteristics. We conducted surveys and interviews with 193 pregnant women and new mothers (clients; 8.1% refusal). Table 1 presents participant characteristics by maternal status and compares the study sample to results from the 2014 Kenya DHS (Kenya National Bureau of Statistics et al., 2015). Nearly a third of the sample consisted of pregnant women in their second or third trimester. The sample resembled the broader population of women in terms of literacy, work, parity, and household headship. The women who participated in this study, however, were somewhat poorer and less educated than other women in the region. See Table A4 in the Online Appendix for additional cross-tabulations by caseness.

Prevalence of Depression. Table 2 displays the diagnostic results by maternal status. The overall prevalence of depression ranges from 5.2 to 14.5 percent depending on which reference standard is used. Defining depression based on DSM-5 criteria produces the most conservative estimate of 5.2 percent. In contrast, asking counselors to base their diagnosis of depression on clinical judgment but leaving this unspecified increases the prevalence estimate to 14.5 percent. Requiring concordance between local counselors and clients’ assessment of their own functioning produces an estimate of 6.2 percent that is more in line with the DSM-5 estimate. Across all definitions, the prevalence of depression by maternal status reflects the overall prevalence. See Table A5 in the Online Appendix to view the pattern of cases and non-cases by case definition.

Item Analysis. Figure 2 displays the extent to which items discriminated between cases and non-cases. This plot shows each item’s gradient score for two different definitions

of caseness: DSM-5 (triangle) and counselor clinical judgment (circle). Items with scores to the right of the solid vertical line at zero were endorsed by a higher percentage of cases compared to non-cases. The larger the gradient score, the bigger the difference in endorsement between cases and non-cases. In addition to having large positive gradient scores, items should have low endorsement by non-cases. Items meeting this criterion have a solid black fill. We adopted the 20 items with a gradient score > 0.05 (based on the local definition) and less than 25 percent endorsement by non-cases (solid black shapes to the right of the dashed vertical line). This includes new items such as `New-19` "feeling hopeless about the future" and PHQ-9 items such as `PHQ-9-3` "trouble falling or staying asleep, or sleeping too much".

Item Selection. Figure A1 in the Online Appendix shows a high density scatterplot of the internal consistency reliability and accuracy of all 616,645 combinations of these 20 screening items in sets of 2 through 10. The best combination is a set of 9 items that includes 2 revised EPDS items, 4 items from other Western psychiatric scales, 2 items generated by the focus groups, and 1 item suggested by the expert panel. We refer to this new scale as the Perinatal Depression Screening, or PDEPS. Item psychometrics are presented in Table 3. Table A6 in the Online Appendix lists the English and Kiswahili translations of the PDEPS items.

Diagnostic Validity. As shown in Table 4, the optimal cutoff for the new PDEPS tool is greater than or equal to 13. Using this cutoff, the PDEPS correctly classifies 90 percent of clients according to the DSM-5 results, with a sensitivity and specificity of 0.90 and 0.90, respectively. A PDEPS score at or above this cutoff is 8.62 times more likely among women diagnosed with depression than women without a diagnosis (DSM-5). See Table A7 in the Online Appendix for estimates of uncertainty.

The PDEPS outperforms the PHQ-9 and both versions of the EPDS in terms of classification accuracy (DSM-5). The optimal cutoff scores for the original EPDS and PHQ-9 are 16 and 15, respectively—higher than what most of the published literature

would recommend (see (Kroenke, Spitzer, & Williams, 2001) and Table A8 in the Online Appendix). While the estimates of sensitivity and specificity are lower for the EPDS and PHQ-9 compared to the PDEPS, these estimates are within the range of what is reported in the literature. Table A8 compares these results to other studies of the EPDS in African countries.

The PDEPS also outperforms the EPDS and PHQ-9 in classifying depression according to the 'local' definition—concordance between counselor clinical judgment and client rating of functioning—but the evidence is more equivocal. The PDEPS correctly classifies more true negatives and is more accurate overall compared to the other scales, but it has a higher false negative rate.

Construct Validity. All scales discriminate between cases and non-cases. As shown in Table 5, the mean PDEPS score among cases is twice as large as the mean score among non-cases. The difference between cases and non-cases is smaller for the other scales, but still statistically significant. Each scale is also negatively correlated with the client and counselor ratings of client functioning, exhibiting the expected inverse relationship between depression severity and functioning. The magnitude of these correlations ranges from -0.24 to -0.38, providing evidence of convergent validity.

Reliability. Table 6 presents estimates of internal consistency and test-retest reliability. Each scale demonstrates acceptable internal consistency reliability as measured by Cronbach's alpha, ranging from 0.77 to 0.83. Estimates of test-retest reliability range from 0.57 to 0.67 across scales for identical administration within 7 days (see Table 3 for estimates of item-level test-retest reliability). Estimates are smaller for retests administered via an automated phone service—from 0.36 to 0.53—but the difference was not statistically significant in all cases.

If women perceived automated phone administration to be more private and confidential, phone retests would be expected to be less reliable as women report more symptoms of depression over the phone. To test this hypothesis, we regressed depression

scores on an indicator of phone retest. As shown in Table A9 in the Online Appendix, mean depression scores are up to 19.9 percent higher among women who were randomly assigned to complete the retest via the automated phone service compared to women who responded to questions read by an enumerator.

Correlates of Depression. Table 7 reports the correlates of the PDEPS score. There is an inverse relationship between household wealth and depression such that women from wealthier households endorsed fewer symptoms of depression. At the same time, however, work is associated with more depression. Living with a spouse or partner is also associated with lower depression scores, but this relationship is not statistically significant.

Discussion

This study demonstrates that the EPDS and PHQ-9 screening tools have acceptable sensitivity and specificity for detecting major depressive episode (DSM-5) among pregnant women and new mothers in Kenya. The EPDS diagnostic validity results are at the low end of what is reported in other studies of African samples, and our recommended cutoff of ≥ 16 is notably higher that what these other studies report (see Tsai et al. 2013 and Table A8 in the Online Appendix), but our results confirm that the EPDS and PHQ-9 are valid instruments for this setting. They may not be not optimal, however.

The new scale we developed through a process of free listing, card sorting, and item analysis—the Perinatal Depression Screening, or PDEPS—performs better on all metrics of classification accuracy with respect to DSM-5 caseness. We recruited too few pregnant women to make strong conclusions about the diagnostic validity of the PDEPS with respect to detecting depression during the antenatal period, but overall classification accuracy was high. Our postpartum sample was twice as large, however, and results suggest that the PDEPS outperforms the EPDS and PHQ-9 in terms of diagnostic accuracy during this period.

We developed the PDEPS through a hybrid emic and etic approach that blended

Western psychiatric concepts with locally derived idioms of distress. This work drew inspiration from several previous studies that took a similar approach to combining Western and indigenous concepts (e.g., Bass, Ryder, Lammers, Mukaba, & Bolton, 2008; Kaaya, Lee, Mbwambo, Smith-Fawzi, & Leshabari, 2008; Nhiwatiwa, Patel, & Acuda, 1998; Patel, Simunyu, Gwanzura, Lewis, & Mann, 1997). In particular, we can compare our results to Bass et al. (2008) who used similar methods in the Democratic Republic of Congo to develop and validate a locally-derived measure of postpartum depression (*Maladi ya Souci*, a syndrome of worry) that contained some EPDS items. The authors found that their 16-item local syndrome scale performed similarly to a shorter version of the EPDS and another Western scale in terms of diagnostic accuracy.

Unlike Bass et al. (2008) who found that their locally-derived scale included all of the diagnostic symptom categories for MDD, we found that the best combination of items on the PDEPS deviates somewhat from DSM-5 criteria. For instance, one of the most discriminating items, "Feeling like you just want to go back to your maternal home", was suggested by the expert panel of Kenyan mental health professionals who are trained from a Western model but whose practice is informed by local customs. This particular custom refers to the tradition that women often leave their maternal home upon marriage and resettle in their husband's village. Wanting to go back to your maternal home would signal difficulty coping with present circumstances. Given the culturally-anchored nature of this item, some adaptation might be required if the PDEPS is to be used in other settings where this behavior is not a custom. Other PDEPS items more clearly reflect the universal nature of depression, such as "feeling hopeless", "feeling anxious or worried for no good reason", and "crying because of sadness"—the latter two overlapping directly with the EPDS.

The study by Bass et al. (2008) is also an interesting comparison because the authors did not rely on standard clinical interviews to assess caseness. Instead, they considered a woman a 'case' if a key informant identified her as suffering from the local syndrome *and* if the woman self-identified as having the syndrome. In addition to using the SCID-5-RV as a

reference criterion, we also separately examined 'local' cases defined by concordance between a counselor's clinical judgment—not bound by DSM-5 criteria—and the woman's self-report of poor functioning. As far as we know, ours is the first study to compare Western and local approaches to define caseness.

Interestingly, both definitions of caseness lead to a similar overall estimate of prevalence: 5.2 percent using a DSM-5 gold-standard interview and 6.2 percent using a local definition based on concordance between counselor clinical judgment and client self-assessment. These results are internally consistent, but represent a divergence from other published prevalence estimates. For instance, a systematic review by Villegas et al. (2011) reported a point estimate of 31.3 percent in developing countries (95%CI 21.3 to 43.5). Another systematic review of common perinatal mental disorders in low- and lower-middle-income countries—so a broader scope than just depressive disorders—estimated an overall prevalence of 18.6 percent (95%CI 18.0 to 19.2; Fisher et al., 2012). A third systematic review limited to studies of African samples—but none conducted in Kenya—reported a prevalence of depression of 11.3 percent during the antenatal period (95%CI 9.5 to 13.1) and 18.3 percent during the postnatal period (95%CI 17.5 to 19.1; Sawyer, Ayers, & Smith, 2010).

If we relax the requirement for concordance and consider only counselor clinical judgment, our estimate of prevalence increases to 14.5 percent. This might be most comparable to the Fisher et al. (2012) estimate of 18.6 percent for common perinatal mental disorders, assuming that counselors in our study might have considered a broader range of symptoms than strictly depression when labeling cases when relying on their clinical judgment. In any case, this higher estimate should probably be considered an upper-bound on prevalence with an understanding that the DSM-5 gold-standard and the local concordance gold-standard produces estimates half as large.

While this study affirms the validity of screening for perinatal depression, it also raises an interesting question about how to implement screening at scale. Low literacy

rates in places like rural Kenya preclude a complete reliance on self-administration. Nurse-administered screenings are of course possible at primary health centers, but this approach is limited by other demands on staff time and the need for additional training. These are barriers that can be overcome, however there is still the issue that facility-based screening will not reach all pregnant women and new mothers, particularly in settings were antenatal care is not universal and where rates of facility delivery remain low. To increase coverage, we should consider how automated screenings delivered via phone calls could help to overcome all of these barriers.

We examined the test-retest reliability of automated phone administration and found that retest surveys conducted via interactive voice response were somewhat less reliable than in-person retest surveys administered by the same enumerator. However, we hypothesized that the private nature of automated phone administration would lead to more endorsement of depression symptoms, thus making the phone retests appear less reliable. This is what we found, and it is consistent with other research on interactive voice response systems. Kobak et al. (1999) found that U.S. patients reported more embarrassment in acknowledging depression symptoms to a live clinician compared to an automated voice system, and Lieberman et al. (2012) reviewed the use of automated voice screening for medical research and concluded that automated interviews give patients a sense of anonymity that leads to increased reporting compared to in-person interviews. More work is needed to understand the potential uses of and barriers to automated screening at scale in low-income countries, particularly for rural health systems.

Phone screenings would also offer the opportunity to assess voice samples. All of the current methods of screening for depression, assessing severity, and monitoring response to treatment rely on either patient-report or clinician judgment, both of which can be subjective and error prone. The search for more objective biomarkers of depression has led researchers to study how depression affects speech. Findings from a recent randomized controlled trial in the U.S. demonstrated that it is feasible to obtain valid measures of

depression severity and response to treatment via the analysis of vocal recordings captured via an automated phone system (Mundt, Vogel, Feltner, & Lenderking, 2012). As part of a Phase 4 randomized trial of treatments for Major Depression, 105 patients assigned to treatment arms provided speech samples in addition to completing several clinical assessments of depression severity and response to treatment. The study found that changes in speech patterns were associated with clinical outcomes, suggesting that there may be clinically important vocal biomarkers of depression. This represents a potentially important new direction for research in wide-scale depression screening in low-income settings.

Limitations

A strength of the current study was the development of a community sampling frame and the use of probability sampling. As demonstrated in the results, the sample of pregnant women and new mothers recruited for this study resembled women in the region in many respects. Despite being representative, however, our sample was recruited from a relatively small catchment area—a 2 kilometer radius around a particular primary health center. The results may generalize to other rural areas in the region, but it is not clear how the findings might apply in different regions or urban areas. Also, while this study addressed the antenatal and postnatal period, the sample size was not large enough to permit a full investigation of differences by period. A lower than expected prevalence rate led to wide confidence intervals on estimates of sensitivity given the sample size. Additionally, it is important to note that the gold-standard SCID-5-RV itself has not been specifically validated in Kenya as far as we know, but the DSM-5 on which it is based is the diagnostic standard for mental health professionals in Kenya. It is also worth noting that participants did not complete the survey and clinical interviews on the same day, but the order of completion was balanced by random assignment.

Conclusions

The EPDS and PHQ-9 are valid and reliable screening tools for perinatal depression in rural Western Kenya, but a new 9-item locally-developed tool called the Perinatal Depression Screening (PDEPS) that blends Western psychiatric concepts and local idioms of distress may be a more useful alternative. At less than 10 percent, the prevalence of depression in this region appears to be lower than other published estimates for African and other low-income countries. Additional research is needed to confirm this finding and to explore how to implement depression screening at scale, potentially through mobile phones and automated voice services. Additional research is also needed to develop and validate instruments for co-morbid symptoms of anxiety during the perinatal period.

References

Abiodun, O. A. (2006). Postnatal depression in primary care populations in nigeria. *General Hospital Psychiatry*, *28*(2), 133–136.

Adewuya, A. O. (2006). Early postpartum mood as a risk factor for postnatal depression in nigerian women. *The American Journal of Psychiatry*, *163*(8), 1435–1437.

Adewuya, A. O., Eegunranti, A. B., & Lawal, A. M. (2005). Prevalence of postnatal depression in western nigerian women: a controlled study. *International Journal of Psychiatry in Clinical Practice*, *9*(1), 60–64.

Adewuya, A. O., Ola, B. A., Dada, A. O., & Fasoto, O. O. (2006). Validation of the edinburgh postnatal depression scale as a screening tool for depression in late pregnancy among nigerian women. *Journal of Psychosomatic Obstetrics and Gynaecology*, *27*(4), 267–272.

Agoub, M., Moussaoui, D., & Battas, O. (2005). Prevalence of postpartum depression in a moroccan sample. *Archives of Women's Mental Health*, *8*(1), 37–43.

Albert L. Siu, & the US Preventive Services Task Force. (2016). Screening for depression in adults: Us preventive services task force recommendation statement. *JAMA*, *315*(4), 380–387.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.). APA.

Bass, J. K., Ryder, R. W., Lammers, M.-C., Mukaba, T. N., & Bolton, P. A. (2008). Post-partum depression in kinshasa, democratic republic of congo: Validation of a concept using a mixed-methods cross-cultural approach. *Tropical Medicine & International Health*, *13*(12), 1534–1542.

Beck, C. T. (1998). The effects of postpartum depression on child development: a meta-analysis. *Archives of Psychiatric Nursing*, *12*(1), 12–20.

Bolton, P. (2001a). Cross-cultural validity and reliability testing of a standard psychiatric assessment instrument without a gold standard. *The Journal of Nervous and Mental*

Disease, *189*(4), 238–242.

Bolton, P. (2001b). Local perceptions of the mental health effects of the rwandan genocide. *The Journal of Nervous and Mental Disease*, *189*(4), 243–248.

Chibanda, D., Mangezi, W., Tshimanga, M., Woelk, G., Rusakaniko, P., Stranix-Chibanda, L., ... Shetty, A. K. (2010). Validation of the edinburgh postnatal depression scale among women in a high hiv prevalence area in urban zimbabwe. *Archives of Women's Mental Health*, *13*(3), 201–206.

Cox, J. L., Holden, J. M., & Sagovsky, R. (1987). Detection of postnatal depression. development of the 10-item edinburgh postnatal depression scale. *The British Journal of Psychiatry*, *150*(6), 782–786.

DHS Program. (n.d.-a). *Dhs model questionnaires.* Retrieved 2016-03-07, from `http://dhsprogram.com/What-We-Do/Survey-Types/DHS-Questionnaires.cfm`

DHS Program. (n.d.-b). *Wealth index construction.* Retrieved 2016-02-01, from `http://www.dhsprogram.com/topics/wealth-index/Wealth-Index-Construction.cfm`

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoSOne*, *10*(4), e0121945.

Field, T., Diego, M., Dieter, J., Hernandez-Reif, M., Schanberg, S., Kuhn, C., ... Bendell, D. (2004). Prenatal depression effects on the fetus and the newborn. *Infant Behavior and Development*, *27*(2), 216–229.

First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. (2015). *Structured clinical interview for DSM-5—research version* (Tech. Rep.). American Psychiatric Association.

Fisher, J., Mello, M. C. d., Patel, V., Rahman, A., Tran, T., Holton, S., & Holmes, W. (2012). Prevalence and determinants of common perinatal mental disorders in women in low-and lower-middle-income countries: a systematic review. *Bulletin of the World Health Organization*, *90*(2), 139–149.

Gelaye, B., Rondon, M. B., Araya, R., & Williams, M. A. (2016). Epidemiology of maternal depression, risk factors, and child outcomes in low-income and middle-income countries. *The Lancet Psychiatry*, *3*(10), 973–982.

Gentile, S. (2017). Untreated depression during pregnancy: Short-and long-term effects in offspring. a systematic review. *Neuroscience*, *342*, 154–166.

Gibson, J., McKenzie-McHarg, K., Shakespeare, J., Price, J., & Gray, R. (2009). A systematic review of studies validating the edinburgh postnatal depression scale in antepartum and postpartum women. *Acta Psychiatrica Scandinavica*, *119*(5), 350–364.

Goldberg, D. P. (1972). *The detection of psychiatric illness by questionnaire: A technique for the identification and assessment of non-psychotic psychiatric illness* (No. 21). Oxford U. Press.

Grigoriadis, S., VonderPorten, E. H., Mamisashvili, L., Tomlinson, G., Dennis, C.-L., Koren, G., ... others (2013). The impact of maternal depression during pregnancy on perinatal outcomes: a systematic review and meta-analysis. *Journal of Clinical Psychiatry*, *74*(4), e321–e341.

Hanlon, C., Medhin, G., Alem, A., Araya, M., Abdulahi, A., Hughes, M., ... Prince, M. (2008). Detecting perinatal common mental disorders in ethiopia: validation of the self-reporting questionnaire and edinburgh postnatal depression scale. *Journal of Affective Disorders*, *108*(3), 251–262.

Joshi, R., Alim, M., Kengne, A. P., Jan, S., Maulik, P. K., Peiris, D., & Patel, A. A. (2014). Task shifting for non-communicable disease management in low and middle income countries–a systematic review. *PloS one*, *9*(8), e103754.

Junge, C., Garthus-Niegel, S., Slinning, K., Polte, C., Simonsen, T. B., & Eberhard-Gran, M. (2017). The impact of perinatal depression on children's social-emotional development: A longitudinal study. *Maternal and Child Health Journal*, *21*(3), 607–615.

Kaaya, S. F., Lee, B., Mbwambo, J. K., Smith-Fawzi, M. C., & Leshabari, M. T. (2008). Detecting depressive disorder with a 19-item local instrument in tanzania. *The International Journal of Social Psychiatry*, *54*(1), 21–33.

Kenya National Bureau of Statistics, Kenya Ministry of Health, National AIDS Control Council, Kenya Medical Research Institute, National Council for Population and Development, & ICF International. (2015). *Kenya demographic and health survey 2014* (Tech. Rep.). KNBS.

Khalifeh, H., Hunt, I. M., Appleby, L., & Howard, L. M. (2016). Suicide in perinatal and non-perinatal women in contact with psychiatric services: 15 year findings from a uk national inquiry. *The Lancet Psychiatry*, *3*(3), 233–242.

Kobak, K. A., Greist, J. H., Jefferson, J. W., Mundt, J. C., & Katzelnick, D. (1999). Computerized assessment of depression and anxiety over the telephone using interactive voice response. *MD computing: computers in medical practice*, *16*(3), 64.

Kohrt, B. A., Jordans, M. J., Tol, W. A., Luitel, N. P., Maharjan, S. M., & Upadhaya, N. (2011). Validation of cross-cultural child mental health and psychosocial research instruments: adapting the depression self-rating scale and child PTSD symptom scale in nepal. *BMC Psychiatry*, *11*(1), 127.

Kroenke, K., Spitzer, R., & Williams, J. (2001). The phq-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606-613.

Lawrie, T. A., Hofmeyr, G. J., de Jager, M., & Berk, M. (1998). Validation of the edinburgh postnatal depression scale on a cohort of south african women. *South African Medical Journal*, *88*(10), 1340–1344.

Lieberman, G., & Naylor, M. R. (2012). Interactive voice response technology for symptom monitoring and as an adjunct to the treatment of chronic pain. *Translational behavioral medicine*, *2*(1), 93–101.

López-Ratón, M., Rodríguez-Álvarez, M. X., Suárez, C. C., & Sampedro, F. G. (2014). OptimalCutpoints: An R package for selecting optimal cutpoints in diagnostic tests.

Journal of Statistical Software, *61*(8), 1–36. Retrieved from http://www.jstatsoft.org/v61/i08/

Monahan, P. O., Shacham, E., Reece, M., Kroenke, K., Ong'or, W. O., Omollo, O., ... Ojwang, C. (2009). Validity/reliability of PHQ-9 and PHQ-2 depression scales among adults living with HIV/AIDS in western kenya. *Journal of General Internal Medicine*, *24*(2), 189–197.

Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lenderking, W. R. (2012). Vocal acoustic biomarkers of depression severity and treatment response. *Biological Psychiatry*, *72*(7), 580–587.

Nhiwatiwa, S., Patel, V., & Acuda, W. (1998). Predicting postnatal mental disorder with a screening questionnaire: a prospective cohort study from zimbabwe. *Journal of Epidemiology and Community Health*, *52*(4), 262–266.

Oates, M. (2003). Perinatal psychiatric disorders: a leading cause of maternal morbidity and mortality. *British Medical Bulletin*, *67*(1), 219–229.

Omoro, S. A. O., Fann, J. R., Weymuller, E. A., Macharia, I. M., & Yueh, B. (2006). Swahili translation and validation of the patient health questionnaire-9 depression scale in the kenyan head and neck cancer patient population. *The International Journal of Psychiatry in Medicine*, *36*(3), 367–381.

Patel, V., Simunyu, E., Gwanzura, F., Lewis, G., & Mann, A. (1997). The shona symptom questionnaire: the development of an indigenous measure of common mental disorders in harare. *Acta Psychiatrica Scandinavica*, *95*(6), 469–475.

R Core Team. (2015). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rahman, A., Hafeez, A., Bilal, R., Sikander, S., Malik, A., Minhas, F., ... Creed, F. (2016). The impact of perinatal depression on exclusive breastfeeding: a cohort study. *Maternal & Child Nutrition*, *12*(3), 452–462.

Rahman, A., Malik, A., Sikander, S., Roberts, C., & Creed, F. (2008). Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural pakistan: a cluster-randomised controlled trial. *The Lancet*, *372*(9642), 902–909.

Rochat TJ, Richter LM, Doll HA, Buthelezi NP, Tomkins A, & Stein A. (2006). Depression among pregnant rural south african women undergoing hiv testing. *JAMA*, *295*(12), 1373–1378.

Sawyer, A., Ayers, S., & Smith, H. (2010). Pre- and postnatal psychological wellbeing in africa: A systematic review. *Journal of Affective Disorders*, *123*(1–3), 17–29.

Surkan, P. J., Patel, S. A., & Rahman, A. (2016). Preventing infant and child morbidity and mortality due to maternal depression. *Best Practice & Research Clinical Obstetrics & Gynaecology*, *36*, 156–168.

Sweetland, A. C., Belkin, G. S., & Verdeli, H. (2014). Measuring depression and anxiety in sub-saharan africa. *Depression and Anxiety*, *31*(3), 223–232.

Taiwo, O. J., & Olayinka, O. O. (2007). The validation of edinburgh postpartum depression scale (epds) in north central nigeria. *Journal of Medicine in the Tropics*, *9*(2), 29–40.

Tesfaye, M., Hanlon, C., Wondimagegn, D., & Alem, A. (2010). Detecting postnatal common mental disorders in addis ababa, ethiopia: validation of the edinburgh postnatal depression scale and kessler scales. *Journal of Affective Disorders*, *122*(1-2), 102–108.

Tsai, A. C., Scott, J. A., Hung, K. J., Zhu, J. Q., Matthews, L. T., Psaros, C., & Tomlinson, M. (2013). Reliability and validity of instruments for assessing perinatal depression in african settings: Systematic review and meta-analysis. *PLoSOne*, *8*(12), e82521.

Uwakwe, R., & Okonkwo, J. E. N. (2003). Affective (depressive) morbidity in puerperal nigerian women: validation of the edinburgh postnatal depression scale. *Acta*

Psychiatrica Scandinavica, *107*(4), 251–259.

Villegas, L., McKay, K., Dennis, C.-L., & Ross, L. E. (2011). Postpartum depression among rural women from developed and developing countries: A systematic review. *The Journal of Rural Health*, *27*(3), 278–288.

Weobong, B., Akpalu, B., Doku, V., Owusu-Agyei, S., Hurt, L., Kirkwood, B., & Prince, M. (2009). The comparative validity of screening scales for postnatal common mental disorder in kintampo, ghana. *Journal of Affective Disorders*, *113*(1-2), 109–117.

Wiesmann, U., Kiteme, B., & Mwangi, Z. (2014). *Socio-economic atlas of kenya: Depicting the national population census by county and sub-location* (Tech. Rep.). KNBS.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413.

Table 1
Participant characteristics by maternal status

Characteristic	Sample			Kenya	
	Pregnant *n*=61	Postpartum *n*=132	All *n*=193	Estimate	Reference
Mean age (SD)	26.6 (5.5)	26.6 (6.0)	27.1 (5.9)		
Completed primary school (%)	47.5	40.9	43.0	58.0	Bungoma County, 2014 KDHS
Literate (%)	83.6	90.2	88.1	88.7	Bungoma County, 2014 KDHS
Married or cohabiting (%)	91.8	87.1	88.6	59.7	National, 2014 KDHS
Mean parity (SD)	3.0 (2.5)	3.8 (2.4)	3.5 (2.5)	3.8	Currently married women, 2014 KDHS
Worked for money past 12 mo (%)	57.4	67.7	64.4	61.7	Western region, 2014 KDHS
Head of household (%)	34.4	34.8	34.7	35.8	Rural, 2014 KDHS
Mean household size (SD)	4.8 (2.5)	5.9 (2.3)	5.5 (2.4)	4.4	Rural, 2014 KDHS
Poorest two wealth quintiles (%)	57.4	64.4	62.2	50.9	Western province, 2008-09 KDHS

Table 2
Diagnostic results by maternal status

Classifications	Pregnant *n*=61	Postpartum *n*=132	All *n*=193
A. Mean functioning, counselor rating (0-100)	80.2 (11.7)	81.6 (11.5)	81.1 (11.5)
B. Mean functioning, client rating (1-10)	6.4 (1.6)	6.2 (1.7)	6.3 (1.6)
C. Poor functioning (<6), client rating (%)	19.0 (31.1)	45.0 (34.1)	64.0 (33.2)
D. Depression, counselor's 'local' diagnosis (%)	10.0 (16.4)	18.0 (13.6)	28.0 (14.5)
E. Depression, local condordance, C and E (%)	3.3	7.6	6.2
F. Depression, DSM-5 (%)	6.6	4.5	5.2

Table 3
PDEPS item statistics

		DSM-5		Local			
Item	Short label	Non-case Mean (SD)	Case Mean (SD)	Non-case Mean (SD)	Case Mean (SD)	Item-Total	Weighted Kappa
New-21	think better if not born	0.7 (1.0)	1.3 (1.3)	0.7 (1.0)	1.4 (1.2)	0.60	0.52
New-19	feeling hopeless about future	0.6 (1.0)	1.6 (1.1)	0.6 (1.0)	1.6 (1.3)	0.50	0.39
New-24	feeling low compared to others	1.0 (1.1)	1.7 (1.3)	1.0 (1.1)	1.8 (1.1)	0.50	0.42
EPDS-R-4	anxious or worried	0.8 (1.0)	1.9 (1.0)	0.9 (1.0)	1.5 (1.0)	0.53	0.28
New-9	problems with mind	0.4 (0.9)	1.4 (0.8)	0.4 (0.8)	1.5 (1.4)	0.52	0.18
EPDS-R-9	crying	0.7 (1.0)	1.5 (0.7)	0.7 (0.9)	1.5 (0.9)	0.62	0.34
New-27	unable to take care of family	0.8 (1.0)	1.5 (1.1)	0.8 (1.0)	1.5 (1.0)	0.47	0.20
New-30	problems with loved one	0.7 (1.0)	1.3 (1.2)	0.6 (0.9)	1.7 (1.2)	0.37	0.42
New-31	want to go back to maternal home	0.6 (0.9)	1.4 (0.8)	0.7 (0.9)	0.9 (1.1)	0.59	0.51

Note. Weighted kappa is reported here as a measure of test-retest reliability at the item-level. Typically the kappa statistic is framed as a measure of inter-rater reliability. For test-retest reliability, an individual woman is the rater at two time points. Since women were responding on an ordinal scale, we calculated weighted kappa to better reflect the degree of agreement or disagreement between time 1 and time 2 (enumerator administration only).

Table 4
Diagnostic validity

		DSM-5						Local					
Scale	Cut	Sen	Spe	Acc	LRP	LRN	AUC	Sen	Spe	Acc	LRP	LRN	AUC
Combined, N=193													
EPDS, Original	≥ 16	0.70	0.72	0.72	2.50	0.42	0.80	0.83	0.73	0.74	3.12	0.23	0.85
EPDS, Revised	≥ 13	0.60	0.73	0.72	2.23	0.55	0.80	0.42	0.72	0.70	1.50	0.81	0.70
PHQ-9	≥ 15	0.70	0.74	0.73	2.65	0.41	0.79	0.75	0.74	0.74	2.93	0.34	0.78
PDEPS	≥ 13	0.90	0.90	0.90	8.62	0.11	0.89	0.58	0.88	0.86	5.00	0.47	0.86
Pregnant, n=61													
EPDS, Original	≥ 16	1.00	0.75	0.77	4.07	0.00	0.98	0.50	0.71	0.70	1.74	0.70	0.56
EPDS, Revised	≥ 13	0.75	0.79	0.79	3.56	0.32	0.87	0.00	0.75	0.72	0.00	1.34	0.58
PHQ-9	≥ 15	0.75	0.75	0.75	3.05	0.33	0.79	1.00	0.75	0.75	3.93	0.00	0.88
PDEPS	≥ 13	0.75	0.96	0.95	21.38	0.26	0.90	0.00	0.92	0.89	0.00	1.09	0.81
Postpartum, n=132													
EPDS, Original	≥ 16	0.50	0.70	0.69	1.69	0.71	0.70	0.90	0.74	0.76	3.51	0.13	0.91
EPDS, Revised	≥ 13	0.50	0.70	0.69	1.69	0.71	0.77	0.50	0.71	0.69	1.73	0.70	0.72
PHQ-9	≥ 15	0.67	0.73	0.73	2.45	0.46	0.78	0.70	0.74	0.74	2.73	0.40	0.77
PDEPS	≥ 13	1.00	0.86	0.87	7.35	0.00	0.90	0.70	0.87	0.85	5.29	0.35	0.86

Note. Sen=sensitivity. Spe=specificity. Acc=accuracy (1-error rate). LRP=likelihood ratio (positive). LRN=likelihood radio (negative). AUC=area under the receiver operating characteristic curve.

Table 5
Construct validity

		Discriminant Validity						Convergent Validity	
	All	DSM-5			Local			r functioning	
Scale	Mean (SD)	Non-case Mean (SD)	Case Mean (SD)	Diff (%)	Non-case Mean (SD)	Case Mean (SD)	Diff (%)	Client Rating	Counselor Rating
EPDS, Original	12.4 (5.8)	12.0 (5.6)	18.9 (6.2)	56.9**	11.9 (5.5)	20.0 (5.8)	68.1***	-0.24***	-0.38***
EPDS, Revised	8.9 (6.5)	8.5 (6.4)	15.1 (4.3)	77.6***	8.6 (6.5)	12.9 (4.8)	50.7*	-0.29***	-0.28***
PHQ-9	10.4 (6.4)	10.0 (6.3)	16.3 (4.1)	62.6***	10.0 (6.3)	16.0 (4.0)	60.4***	-0.27***	-0.25***
PDEPS	6.7 (5.3)	6.3 (5.2)	13.6 (3.0)	115.8***	6.2 (5.1)	13.3 (4.2)	113.7***	-0.25***	-0.32***

Note. * $p<0.05$. ** $p<0.01$. *** $p<0.001$.

Table 6
Internal consistency and test-retest reliability

	Internal Consistency		Test-retest						
			Enumerator		Automated phone		Difference		
Scale	*N*	Alpha	*n*	*r*	*n*	*r*	*r*	L95%	U95%
EPDS, Original	193	0.78	95	0.57	92	0.53	0.04	-0.17	0.24
EPDS, Revised	193	0.83	95	0.64	92	0.44	0.21	0.00	0.42
PDEPS	193	0.77	95	0.67	92	0.51	0.16	-0.03	0.35
PHQ-9	193	0.81	95	0.62	92	0.36	0.26	0.04	0.49

Note. Women were randomized to complete a re-test survey with the same enumerator who read the questions and recorded responses (same method as the first administration) or via an automated phone screening.

Table 7
Correlates of depression severity

	Dependent variable:
	PDEPS score
Wealth index value	-2.7^{*} (1.6)
Age	-0.01 (0.1)
Working (0/1)	2.3^{**} (0.9)
Currently married or living with a partner (0/1)	-1.3 (1.3)
Parity ≥ 1(0/1)	0.6 (2.0)
Pregnant (0/1)	1.2 (0.9)
Years education completed	-0.1 (0.2)
Literate (0/1)	-0.7 (1.3)
Constant	4.8 (3.3)
Observations	193
R^2	0.1
Adjusted R^2	0.05
Residual Std. Error	5.2 (df = 184)
F Statistic	2.3^{**} (df = 8; 184)
Note:	*p<0.1; **p<0.05; ***p<0.01

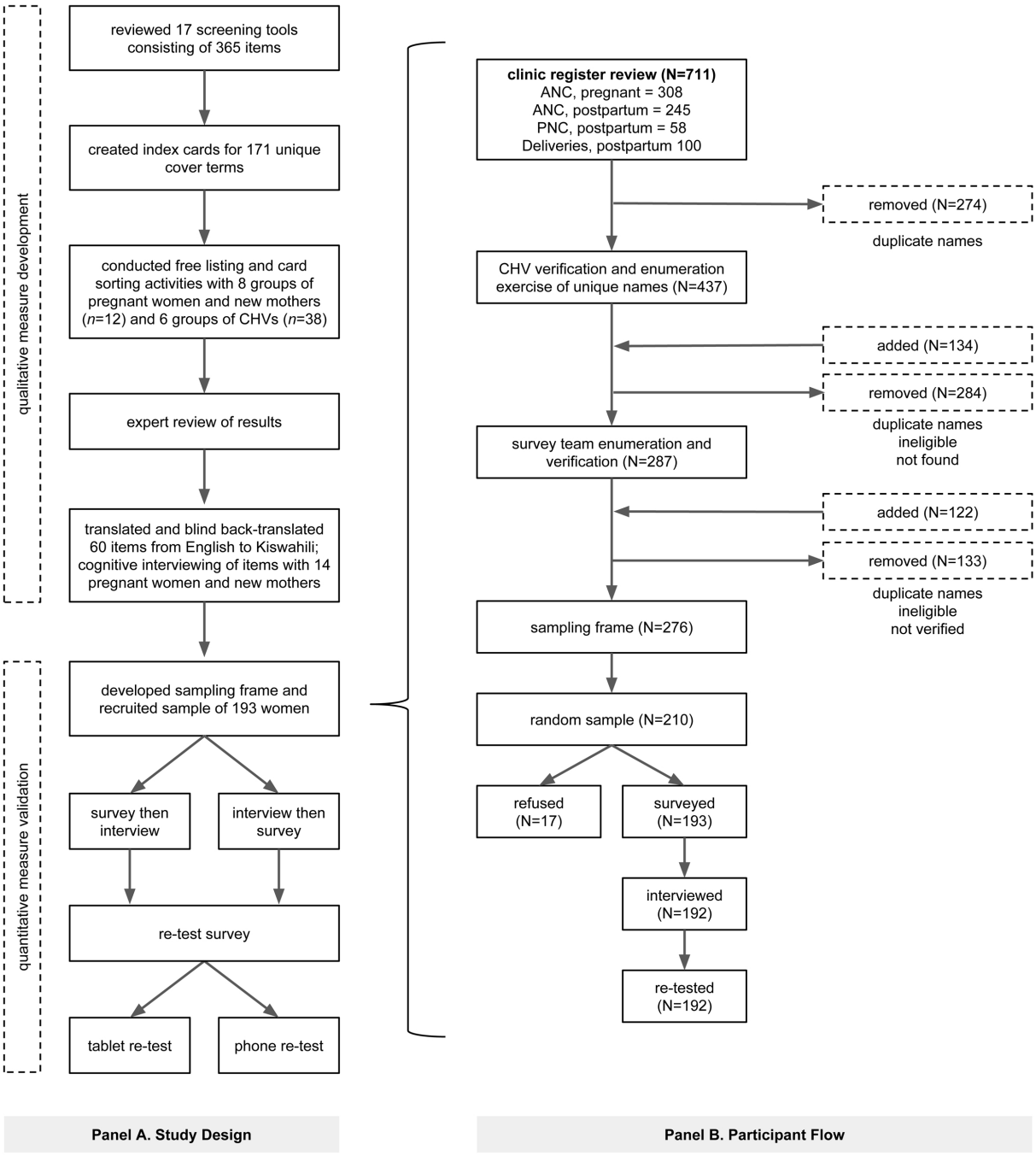

Figure 1. Panel A: Study design and sequence. Panel B. Participant flow diagram for the diagnostic accuracy study.

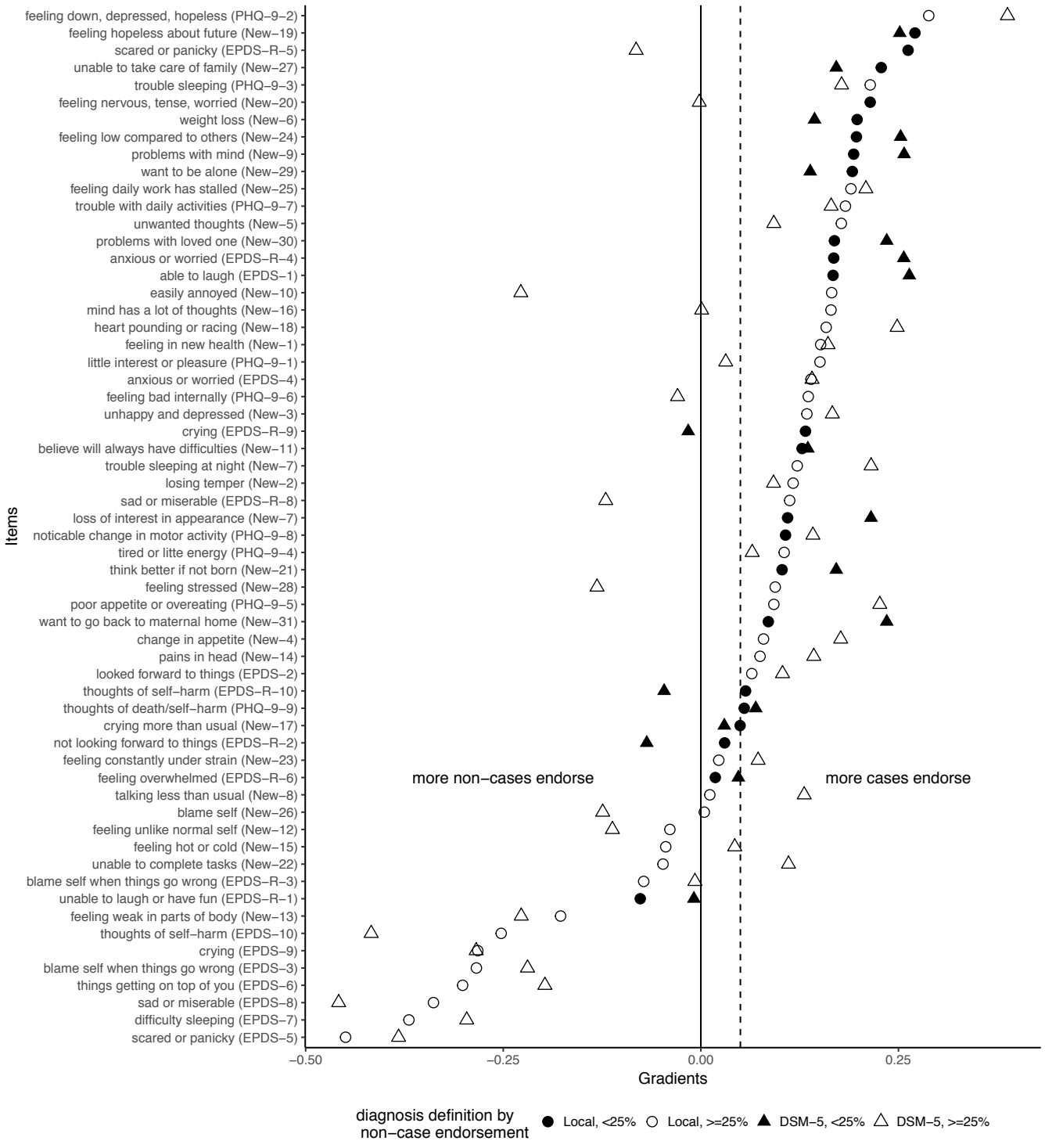

Figure 2. Item gradients showing discrimination between cases and non-cases, sorted by counselor clinical judgment case definition.

Appendix A

Supplemental Tables and Figures

Table A1
Measures of depression reviewed

Measures	Items
SCL-90	90
General Health Questionnaire	60
Pregnancy Risk Assessment Monitoring System Core	38
Response Inventory for Stressful Life Events	36
Hopkins Symptoms Checklist for Depression	24
Pitt Depression Scale	24
Becks Depression and Anxiety Scales	21
Center for Epidemiological Studies Depression Scale	20
Zung's Self Rating Depression Scale	20
Self Reporting Questionnaire	20
Bromley Postnatal Depression Scale	17
Pregnancy Risk Assessment Monitoring System Standard	17
Hospital Anxiety and Depression Scale	14
Edinburgh Postnatal Depression Scale	10
Kessler Mental Health Distress Scale	10
Patient Health Questionnaire-9	9
Pregnancy Risk Assessment Monitoring System 6	6

Note. We removed items included in the "Core" and "Standard" versions of the Pregnancy Risk Assessment Monitoring System (PRAMS) and the Bromley Postnatal Depression Scale (BPDS) because the scales assess a patient's history rather than her depressive symptoms. For instance, the BPDS includes items such as "Did you talk to a psychiatrist because you felt depressed during the first year after the baby was born?" An example PRAMS "Core" item is "Thinking back just before you got pregnant with your new baby, how did you feel about becoming pregnant?".

Table A2
Modifications to the SCID-5-RV

Module	Modification
overview	The sections on demographic data and alcohol and drug use were reduced.
overview	Added: "Imagine a ladder with 10 steps. On the bottom step, Step 1, are women who are really struggling and not doing well. On the top step, Step 10, are women who are doing very well. On which step do you stand?"
A	The following MDE specifiers were not included: with mixed features, with catatonia, with melancholic features, and with atypical features
A	Past MDE not assessed
A	Other mood disorders not assessed
A	Raters could not rule out depression due to another medical condition or a substance; rather they were prompted to investigate and flag the case for discussion
D	Other mood disorders not assessed
D	For MDD, Criterion E not assessed
D	MDD type not assessed
D	MDD seasonal pattern not assessed
D	MDD congruency of psychotic symptoms not assessed
D	MDD with panic attacks not assessed
D	Raters could not rule out depression due to another medical condition or a substance; rather they were prompted to investigate and flag the case for discussion

Note. MDE = Major Depressive Episode. MDD = Major Depressive Disorder.

Table A3
EPDS revisions

Item	Original	Revised
1	I have been able to laugh and see the funny side of things	...feeling unable to laugh or have fun
2	I have looked forward with enjoyment to things	...not looking forward to things
3	I have blamed myself unnecessarily when things went wrong	...blaming yourself when things go wrong
4	I have been anxious or worried for no good reason	...feeling anxious or worried for no good reason
5	I have felt scared or panicky for no very good reason	...feeling scared or panicky for no good reason
6	Things have been getting on top of me	...feeling overwhelmed
7	I have been so unhappy that I have had difficulty sleeping	...trouble sleeping at night
8	I have felt sad or miserable	...feeling sad or miserable
9	I have been so unhappy that I have been crying	...crying because of sadness
10	Thoughts of harming myself have occurred to me	...thoughts of harming yourself

Note. In addition to including the original 10 EPDS items, we also included revised EPDS items that matched the format and response options of the PHQ-9. First, we rephrased the new set of EPDS items so that women could indicate how often they have been bothered by the particular problem in the past two weeks, rather than the past one week. Second, we reworded the three 'positive' EPDS items to express problems.

Table A4
Participant characteristics by maternal status and caseness

	Non-cases			Cases			
Case Definition/Characteristic	Pregnant	Postpartum	Both	Pregnant	Postpartum	Both	All
DSM-5 Diagnosis†							
N	57	125	182	4	6	10	192
Mean age (SD)	26.3 (5.4)	27.4 (5.8)	27.1 (5.7)	30.8 (6.2)	24.5 (8.7)	27.0 (8.1)	27.1 (5.9)
Completed primary school (%)	50.9	40.0	43.4	0.0	50.0	30.0	43.0
Literate (%)	84.2	90.4	88.5	75.0	83.3	80.0	88.1
Married or cohabiting (%)	91.2	87.2	88.5	100.0	83.3	90.0	88.6
Mean parity (SD)	2.9 (2.5)	3.7 (2.3)	3.5 (2.4)	4.2 (1.5)	5.2 (4.7)	4.8 (3.6)	3.5 (2.5)
Worked for money past 12 mo (%)	54.4	66.7	62.8	100.0	100.0	100.0	64.4
Head of household (%)	33.3	32.8	33.0	50.0	66.7	60.0	34.7
Mean household size (SD)	4.8 (2.6)	5.8 (2.2)	5.5 (2.3)	5.8 (1.0)	7.2 (4.3)	6.6 (3.3)	5.5 (2.4)
Poorest wealth quintiles (%)	56.1	64.0	61.5	75.0	83.3	80.0	62.2
Local, Counselor-Client Concordance							
N	59	121	180	2	10	12	192
Mean age (SD)	26.4 (5.4)	27.3 (5.9)	27.0 (5.7)	33.0 (4.2)	28.0 (7.2)	28.8 (6.9)	27.1 (5.9)
Completed primary school (%)	49.2	40.5	43.3	0.0	40.0	33.3	43.0
Literate (%)	83.1	90.1	87.8	100.0	90.0	91.7	88.1
Married or cohabiting (%)	91.5	87.6	88.9	100.0	80.0	83.3	88.6
Mean parity (SD)	2.9 (2.5)	3.6 (2.4)	3.4 (2.5)	5.5 (0.7)	5.4 (2.5)	5.4 (2.3)	3.5 (2.5)
Worked for money past 12 mo (%)	55.9	67.2	63.5	100.0	80.0	83.3	64.4
Head of household (%)	35.6	33.1	33.9	0.0	50.0	41.7	34.7
Mean household size (SD)	4.7 (2.5)	5.8 (2.2)	5.4 (2.4)	7.5 (0.7)	7.3 (2.8)	7.3 (2.5)	5.5 (2.4)
Poorest wealth quintiles (%)	55.9	63.6	61.1	100.0	80.0	83.3	62.2
Local, Counselor Only							
N	51	113	164	10	18	28	192
Mean age (SD)	25.7 (4.8)	27.0 (5.7)	26.6 (5.4)	30.9 (6.9)	29.6 (7.5)	30.0 (7.2)	27.1 (5.9)
Completed primary school (%)	51.0	40.7	43.9	30.0	38.9	35.7	43.0
Literate (%)	84.3	90.3	88.4	80.0	88.9	85.7	88.1
Married or cohabiting (%)	90.2	87.6	88.4	100.0	83.3	89.3	88.6
Mean parity (SD)	2.7 (2.3)	3.5 (2.2)	3.2 (2.2)	4.6 (3.1)	5.6 (3.2)	5.2 (3.2)	3.5 (2.5)
Worked for money past 12 mo (%)	52.9	69.4	64.2	80.0	61.1	67.9	64.4
Head of household (%)	35.3	31.9	32.9	30.0	50.0	42.9	34.7
Mean household size (SD)	4.5 (2.2)	5.6 (2.1)	5.3 (2.2)	6.3 (3.2)	7.3 (2.8)	7.0 (2.9)	5.5 (2.4)
Poorest wealth quintiles (%)	56.9	61.9	60.4	60.0	83.3	75.0	62.2

Note. † Major Depressive Episode or Other specified depressive disorder.

Table A5
Pattern of diagnostic results

Definition	Count	%	Cumlative %
Meets all 3 definitions	3	1.6	1.6
DSM-5 & counselor	5	2.6	4.2
DSM-5 & client	1	0.5	4.7
Client & counselor	9	4.7	9.5
DSM-5 only	1	0.5	10.0
Counselor only	11	5.8	15.8
Client only	51	26.8	42.6
None	109	57.4	100.0

Table A6
Perinatal Depression Screening (PDEPS)

Item	Source	English	Kiswahili
New-9	Western	Thinking that there are problems with your mind	Kufikiria kuwa kuna matatizo na akili yako
New-19	Western	Feeling hopeless about the future	Kuhisi hauna matumaini ya mbeleni
New-21	Western	Thinking it would be better if you had never been born	Kufikiria kuwa ingekuwa bora ikiwa haungezaliwa
New-24	Western	Feeling that you are low when compared to other people	Kuhisi kuwa upo chini ukijilinganisha na watu wengine
New-27	Local	Feeling unable to take care of your children or family	Kuhisi hauna uwezo wa kulinda watoto wako au familia yako
New-30	Local	Having problems with partner or other loved one	Kuwa na shida na mpenzi wako au wengine uwapendao
New-31	Local	Feeling like you just want to go back to your maternal home?	Kuhisi kuwa unataka tu kurudi nyumbani kwa wazazi waliokuzaa
EPDS-R-4	EPDS-R	Feeling anxious or worried for no good reason	Kuhisi wasiwasi bila ya sababu nzuri
EPDS-R-9	EPDS-R	Crying because of sadness	Kulia kwa sababu ya huzuni

Note. "Western" indicates that the item came from an existing screening tool. "Local" indicates that the item was generated by one or more of the focus group discussions.

Table A7
Uncertainty in diagnostic validity

		DSM-5						Local					
		Sensitivity			Specificity			Sensitivity			Specificity		
Scale	Cut	Est	L95	U95	Est	L95	U95	Est	L95	U95	Est	L95	U95
Combined, N=193													
EPDS, Original	≥ 16	0.70	0.35	0.93	0.72	0.65	0.78	0.75	0.43	0.95	0.78	0.72	0.84
EPDS, Revised	≥ 13	0.60	0.26	0.88	0.73	0.66	0.79	0.67	0.35	0.90	0.68	0.61	0.75
PHQ-9	≥ 15	0.70	0.35	0.93	0.74	0.67	0.80	0.75	0.43	0.95	0.74	0.67	0.81
PDEPS	≥ 13	0.90	0.55	1.00	0.90	0.84	0.94	0.75	0.43	0.95	0.78	0.71	0.84
Pregnant, n=61													
EPDS, Original	≥ 16	1.00	0.40	NaN	0.91	0.81	0.97	0.50	0.01	0.99	0.47	0.34	0.61
EPDS, Revised	≥ 13	0.75	0.19	0.99	0.77	0.64	0.87	0.50	0.01	0.99	0.56	0.42	0.69
PHQ-9	≥ 15	0.75	0.19	0.99	0.75	0.62	0.86	1.00	0.16	NaN	0.75	0.62	0.85
PDEPS	≥ 13	0.75	0.19	0.99	0.79	0.66	0.89	0.50	0.01	0.99	0.76	0.63	0.86
Postpartum, n=132													
EPDS, Original	≥ 16	0.67	0.22	0.96	0.63	0.54	0.72	0.80	0.44	0.97	0.79	0.70	0.85
EPDS, Revised	≥ 13	0.50	0.12	0.88	0.70	0.62	0.78	0.80	0.44	0.97	0.66	0.57	0.74
PHQ-9	≥ 15	0.67	0.22	0.96	0.73	0.64	0.80	0.70	0.35	0.93	0.74	0.66	0.82
PDEPS	≥ 13	0.83	0.36	1.00	0.87	0.80	0.93	0.80	0.44	0.97	0.77	0.68	0.84

Note. 95 percent confidence intervals reported.

Table A8
Comparison of EPDS diagnostic validity results across African samples

Study	Country	Sample	Criterion	Period	Case-Control	Uniform	Blinding	Cutoff	Sen	Spe
CURRENT STUDY	Kenya	Community	SCID-5-RV	Antenatal/postnatal (4-36 wk)	No	Yes	Yes	≥ 16	0.70	0.72
Uwakwe & Okonkwo (2003)	Nigeria	Clinical	Clinical interview	Postnatal (unclear)	No	No	Unclear	≥ 9	0.75	0.97
Adewuya et al. (2006)	Nigeria	Clinical	MINI	Antenatal	Yes	No	Yes	≥ 12	1.00	0.96
Agoub et al. (2005)	Morocco	Clinical	MINI	Postnatal (6 mo)	No	No	No	≥ 12	0.92	0.96
Adewuya et al. (2005)	Nigeria	Clinical	SCID	Postnatal (6 wk)	Yes	No	Yes	≥ 9	1.00	0.89
Adewuya (2006)	Nigeria	Clinical	SADS	Postnatal (8 wk)	No	Unclear	Unclear	≥ 9	0.90	0.89
Chibanda et al. (2010)	Zimbabwe	Clinical	Clinical interview	Postnatal (6 wk)	No	Yes	Yes	≥ 12	0.88	0.89
Abiodun (2006)	Nigeria	Clinical	PSE	Postnatal (6 wk)	Yes	No	Yes	≥ 9	0.88	0.84
Rochat et al. (2006)	South Africa	Clinical	SCID	Antenatal	No	Yes	Unclear	≥ 13	0.69	0.78
Tesfaye et al. (2010)	Ethiopia	Clinical	Clinical interview	Postnatal (unclear)	No	Yes	Yes	≥ 8	0.85	0.77
Weobong et al. (2009)	Ghana	Community	CPRS	Postnatal (5-11 wk)	Yes	Yes	Yes	≥ 11	0.78	0.73
Lawrie et al. (1998)	South Africa	Clinical	Clinical interview	Postnatal (6 wk)	No	No	Yes	≥ 12	1.00	0.68
Taiwo & Olayinka (2007)	Nigeria	Clinical	Clinical interview	Postnatal (6 wk)	Yes	No	Yes	≥ 7	0.72	0.62
Hanlon et al. (2008)	Ethiopia	Community	CPRS	Postnatal (unclear)	No	Yes	Unclear	≥ 6	0.77	0.36

Note. Table reproduced in part from Tsai et al. (2013). CPRS = Comprehensive Psychopathological Rating Scale. MINI = Mini International Neuropsychiatric Interview. PSE = Present State Examination. SADS = Schedule for Affective Disorders and Schizophrenia. SCID = Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders. Case-control study design: reference criterion is established in a subset of participants based on the results of the index test. Non-uniform test: index test is not administered in a uniform fashion. Blinded: reference criterion is administered and/or assessed without knowledge of the index test results. Sen = Sensitivity. Spe = Specificity. Weobong et al. (2009) also assessed the diagnostic validity of the PHQ-9 among new mothers in Ghana and found that a cutoff of ≥ 5 had a sensitivity and specificity of 0.94 and 0.75, respectively.

Table A9
Regression of depression severity score on indicator of automated phone administration

	Enumerator	Automated phone				
Scale	Mean	Est	StdErr	% Diff	t	p
EPDS, Original	12.6	2.4	0.85	18.7	2.8	0.01
EPDS, Revised	10.2	0.1	1.03	0.6	0.1	0.95
PDEPS	7.3	1.4	0.91	19.9	1.6	0.11
PHQ-9	9.8	1.7	0.90	17.0	1.9	0.06

Note. Women were randomized to complete a re-test survey with the same enumerator who read the questions and recorded responses (same method as the first administration) or via an automated phone screening.

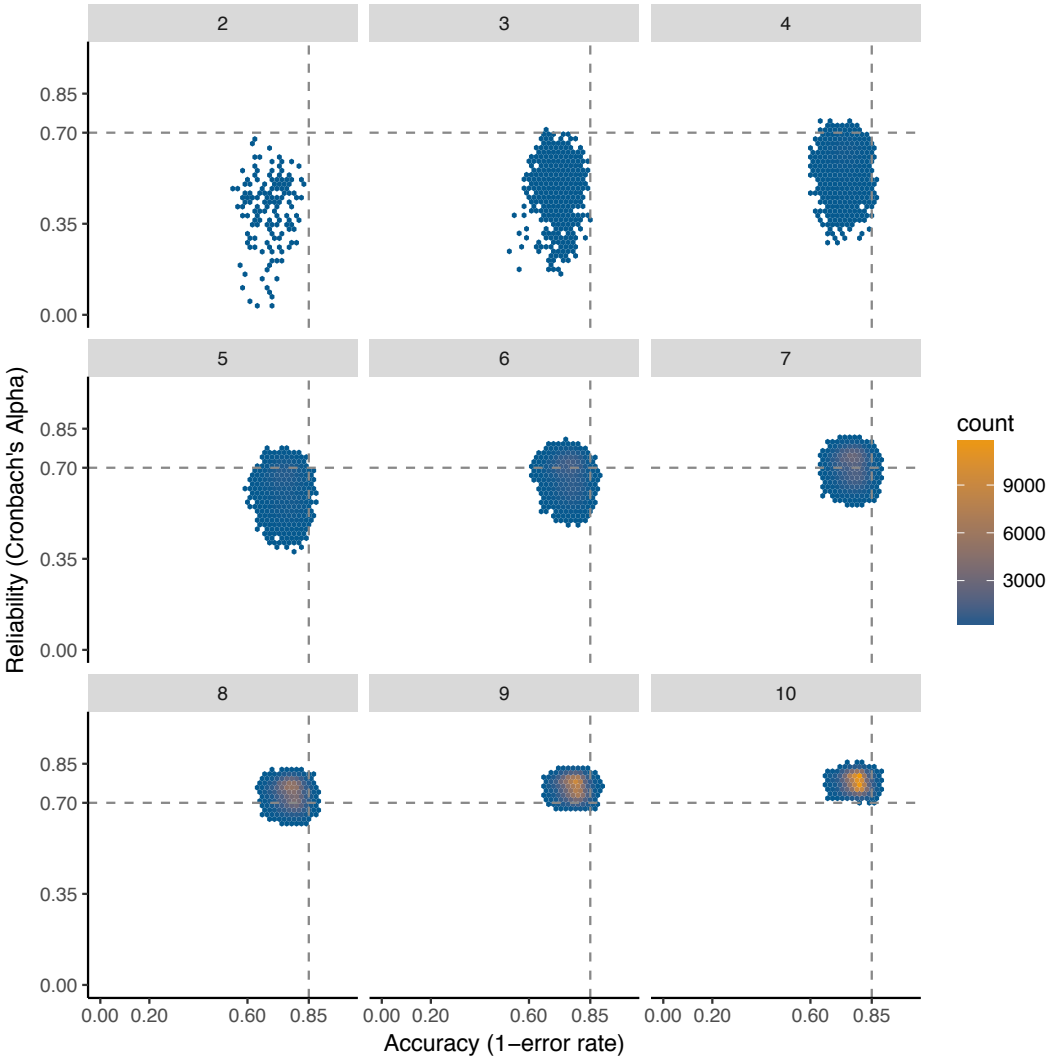

Figure A1. High density scatterplot of the internal consistency reliability and accuracy of all 616,645 combinations of 20 screening items in sets of 2 through 10. DSM-5 definition of depression.

Appendix B

Rationale for Sample Size

Sample Size Justification Submitted to IRBs Prior to Study Launch

In terms of test accuracy, sensitivity is a measure of a test's ability to identify true positive cases of a condition, and specificity, on the other hand, is a measure of true negatives. There is often a tradeoff between sensitivity and specificity. A test with high specificity would result in few false positives, but could misclassify too many patients as negative for the condition when they are really positive (false negative). In the case of life threatening conditions, there is a large cost to false negatives. In the case of perinatal depression, however, we see more harm stemming from false positives than from false negatives. Aside from placing unnecessary burden on overburdened healthcare systems, false positives can also be stigmatizing, especially in a setting like rural Kenya. Thus while aiming for an optimal balance, we have a slight preference for high specificity over high sensitivity.

As reported in the Sweetland et al. (2014) systematic review of 65 depression screening validity studies from 16 countries in sub-Saharan Africa, there is substantial variability in assessment measures, gold standards, and diagnostic accuracy used to assess the criterion validity of depression screening instruments. In Table 1, we summarize reported sensitivity and specificity collapsed across scales, gold standards, and samples.

Summary statistics for reported sensitivity and specificity of depression screening instruments reviewed in Sweetland et al. (2014)

			Average (SD)		Number of Studies	
Scale	Gold Standard	Sample	Sensitivity	Specificity	All	Clinical Interview
Any	Clinical interview	Any	76.4 (17.1)	80.6 (19.3)	16	16
Any	Any	Perinatal	78.0 (10.1)	76.9 (15.6)	17	4
EPDS	Any	Perinatal	82.8 (8.4)	79.1 (19.8)	8	2
PHQ-9	Any	Any	88.3 (4.8)	91.1 (13.9)	3	0

Based on these results, a reasonable expectation for specificity of our new test would be 0.80. Using the method of Chu and Cole (2007), we estimate that we would need to include 130 non-cases to ensure with 0.95 probability that the lower 95% confidence limit does not fall below 0.70, assuming power of 0.80. If the prevalence of depression in this setting is 31.3%, this would mean recruiting 60 cases for a total of 190 women.[1]

```
cases = noncases/[(1-prevalance)/prevalence]
```

Without published estimates of perinatal depression in this part of Kenya to guide us, we have to look elsewhere in the literature. Villegas et al. (2011) found a combined prevalence rate of postpartum depression of 31.3% (95% CI 21.3-43.5%) across 10

[1] If our estimate of specificity is higher as hoped, say 0.95, then we would only need to recruit 68 non-cases (31 cases for a representative sample) to ensure that the lower 95% confidence limit does not fall below 0.85 with 0.95 probability.

studies conducted in developing countries.[2]

If the sensitivity of our new test is also 0.80—a reasonable estimate from Table 1—and if we successfully recruit 60 cases, then we can ensure with 0.95 probability that the lower 95% confidence limit will not fall below 0.64. Raising this lower limit to 0.70 would require us to recruit 130 cases, and that that is not feasible given our current resources; however, if sensitivity is higher as hoped—for instance, 0.85, 0.90, or 0.95—then 60 cases would be sufficient to ensure that the lower 95% confidence limit would be 0.70, 0.75, or 0.80, respectively.

Therefore, we conclude that an optimal validation sample size is 190 women, including 60 cases and 130 non-cases. If there is attrition between the survey and the clinical assessment, we will have to recruit additional women. Attrition is unlikely to exceed 10%, so we anticipate that the maximum required validation sample would be 209.[3]

[2] If the "true" prevalence is closer to the lower confidence limit estimated by Villegas et al. (2011), then we would only need to recruit 36 cases—not 60—for a total sample of 166 women.

[3] It is possible that we could exceed this total if we struggle to find cases and inadvertently oversample non-cases in an effort to recruit 60 cases.

Appendix C

STARD Checklist

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	abstract
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	yes
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	4-5
	4	Study objectives and hypotheses	5
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	5; prospective
Participants	**6**	Eligibility criteria	5-6
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	5-6
	8	Where and when potentially eligible participants were identified (setting, location and dates)	5-6
	9	Whether participants formed a consecutive, random or convenience series	11-12; random
Test methods	**10a**	Index test, in sufficient detail to allow replication	6-7; EPDS, PHQ9, local
	10b	Reference standard, in sufficient detail to allow replication	7-8; SCID-5-RV and local
	11	Rationale for choosing the reference standard (if alternatives exist)	7-8
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	14; optimal
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	7-8; pre-specified
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	12; blinded
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	12; blinded
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	NA
	15	How indeterminate index test or reference standard results were handled	13
	16	How missing data on the index test and reference standard were handled	15; 1 incomplete case dropped
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	15
	18	Intended sample size and how it was determined	11-12
RESULTS			
Participants	**19**	Flow of participants, using a diagram	F.1
	20	Baseline demographic and clinical characteristics of participants	T.1
	21a	Distribution of severity of disease in those with the target condition	NA
	21b	Distribution of alternative diagnoses in those without the target condition	NA
	22	Time interval and any clinical interventions between index test and reference standard	12
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	T4
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	T.4 and T.A7
	25	Any adverse events from performing the index test or the reference standard	NA
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	23
	27	Implications for practice, including the intended use and clinical role of the index test	22-24
OTHER INFORMATION			
	28	Registration number and name of registry	NA
	29	Where the full study protocol can be accessed	NA
	30	Sources of funding and other support; role of funders	Duke

References

Abiodun, O. A. (2006). Postnatal depression in primary care populations in Nigeria. *General Hospital Psychiatry, 28(2)*, 133–136.

Adewuya, A. O. (2006). Early postpartum mood as a risk factor for postnatal depression in Nigerian women. *The American Journal of Psychiatry, 163(8)*, 1435–1437.

Adewuya, A. O., Eegunranti, A. B., & Lawal, A. M. (2005). Prevalence of postnatal depression in Western Nigerian women: a controlled study. *International Journal of Psychiatry in Clinical Practice, 9(1)*, 60–64.

Adewuya, A. O., Ola, B. A., Dada, A. O., & Fasoto, O. O. (2006). Validation of the Edinburgh Postnatal Depression Scale as a screening tool for depression in late pregnancy among Nigerian women. *Journal of Psychosomatic Obstetrics and Gynaecology, 27(4)*, 267–272.

Agoub, M., Moussaoui, D., & Battas, O. (2005). Prevalence of postpartum depression in a Moroccan sample. *Archives of Women's Mental Health, 8(1)*, 37–43.

Chibanda, D., Mangezi, W., Tshimanga, M., Woelk, G., Rusakaniko, P., Stranix-Chibanda, L., & Shetty, A. K. (2010). Validation of the Edinburgh Postnatal Depression Scale among women in a high HIV prevalence area in urban Zimbabwe. *Archives of Women's Mental Health, 13(3)*, 201–206.

Chu, H., & Cole, S. R. (2007). Sample size calculation using exact methods in diagnostic test studies. *Journal of Clinical Epidemiology, 60(11)*, 1201–1202.

Hanlon, C., Medhin, G., Alem, A., Araya, M., Abdulahi, A., Hughes, M., & Prince, M. (2008). Detecting perinatal common mental disorders in Ethiopia: validation of the self-reporting questionnaire and Edinburgh Postnatal Depression Scale. *Journal of Affective Disorders, 108(3)*, 251–262.

Lawrie, T. A., Hofmeyr, G. J., de Jager, M., & Berk, M. (1998). Validation of the

Edinburgh Postnatal Depression Scale on a cohort of South African women. *South African Medical Journal, 88(10)*, 1340–1344.

Rochat TJ, Richter LM, Doll HA, Buthelezi NP, Tomkins A, & Stein A. (2006). Depression among pregnant rural south african women undergoing hiv testing. *JAMA, 295(12)*, 1373–1378.

Sweetland, A. C., Belkin, G. S., & Verdeli, H. (2014). Measuring depression and anxiety in sub-Saharan Africa. *Depression and Anxiety, 31(3)*, 223–232.

Taiwo, O. J., & Olayinka, O. O. (2007). The validation of Edinburgh Postpartum Depression Scale (EPDS) in North Central Nigeria. *Journal of Medicine in the Tropics, 9(2)*, 29–40.

Tesfaye, M., Hanlon, C., Wondimagegn, D., & Alem, A. (2010). Detecting postnatal common mental disorders in Addis Ababa, Ethiopia: validation of the Edinburgh Postnatal Depression Scale and Kessler Scales. *Journal of Affective Disorders, 122(1–2)*, 102–108.

Uwakwe, R., & Okonkwo, J. E. N. (2003). Affective (depressive) morbidity in puerperal Nigerian women: validation of the Edinburgh Postnatal Depression Scale. *Acta Psychiatrica Scandinavica, 107(4)*, 251–259.

Villegas, L., McKay, K., Dennis, C.-L., & Ross, L. E. (2011). Postpartum depression among rural women from developed and developing countries: A systematic review. *The Journal of Rural Health, 27(3)*, 278–288.

Weobong, B., Akpalu, B., Doku, V., Owusu-Agyei, S., Hurt, L., Kirkwood, B., & Prince, M. (2009). The comparative validity of screening scales for postnatal common mental disorder in Kintampo, Ghana. *Journal of Affective Disorders, 113(1–2)*, 109–117.