Published in final edited form as:

J Eval Clin Pract. 2020 October ; 26(5): 1425–1435. doi:10.1111/jep.13327.

Adaptation and validation of a structured version of the Hamilton Depression Rating Scale for use by non-clinicians in South Africa (AFFIRM-HDRS)

Thandi Davies,

Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town, South Africa.

Emily C Garman,

Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town, South Africa.

Crick Lund,

Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town, South Africa, and the Centre for Global Mental Health, Health Service and Population Research Department, Institute of Psychiatry, Psychology and Neuroscience, King's College London.

Marguerite Schneider

Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town, South Africa.,

Abstract

Rationale, aims and objectives: The HDRS is seen as an international gold standard for the measurement of depression. High rates of untreated depression, including perinatal depression, support the rationale to adapt and validate a structured version of the HDRS to be used by non-clinicians in low-income settings.

Methods: Using previous structured versions of the HDRS as a foundation, the tool was adapted, translated, and then validated with an isi-Xhosa speaking perinatal population in South Africa (n=187), using cognitive testing, test-retest reliability and inter-rater assessments. The AFFIRM-HDRS was compared with the Edinburgh Postnatal Depression Scale (EPDS) using the non-parametric Spearman Rho test to assess concurrent validity. Internal consistency was examined using Cronbach's Alpha, and inter-rater and test-retest reliability were assessed with the intra-class coefficient (ICC). Cohen's Kappa was used to assess the overall percentage agreement for each individual item of the AFFIRM-HDRS.

(corresponding author). Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town, 46 Sawkins Road, Rondebosch, 7700, South Africa. thandi.davies@gmail.com.

Author contributions

TD participated in study design, data collection, analysis and article preparation. EB and MS participated in study design, analysis and article preparation. CL participated in study design and article preparation. All authors have approved the final article.

Conflict of interest

The authors declare that they have no conflicting interests in this publication.

Results: The AFFIRM-HDRS showed good construct and content validity, had significant associations with the EPDS ($Rho=.60$ and $.43$, $p<.001$), and acceptable internal consistency (Cronbach's $\alpha = .74$). Inter-rater reliability and test-retest scores were excellent, with intraclass correlations ranging from $.97$ ($.94 - .99$) to $.98$ ($.97 - .99$) between raters, and test-retest reliability being $.90$ (95% CI: $.86-.93$). The tool performed similarly to previous structured versions. Individual item-rest correlations suggest that the items 'Weight loss or gain', 'Insight', and 'Libido' did not fit well with the overall instrument, but that the rest of the items performed well.

Conclusion: The AFFIRM-HDRS is adequately structured to be used by non-clinicians in an isiXhosa speaking perinatal population. It is hoped that this structured tool can be used to assist with identification and referral of these at-risk populations by non-clinicians in resource constrained environments, thereby playing a role in addressing the treatment gap for perinatal depression in LMICs.

Keywords

Hamilton Depression Rating Scale (HDRS); Adaptation; Reliability; Validity; Depression; LMIC; task sharing; perinatal; South Africa

1. INTRODUCTION

Major depressive disorder is now one of the largest single contributors to the burden of disease globally.¹ Within this, perinatal depression prevalence is high, particularly in low and middle-income countries (LMICs), ranging between 15.6% and 19.8% for pre- and post-natal depression, respectively.² However, there is a wide treatment gap between those requiring services and those receiving them.³ The effect of untreated depression, particularly maternal depression, has numerous negative consequences, such as reduced economic capabilities and productivity, poor care of self and family, and poor infant outcomes.^{4, 5}

With the increase in development of task-sharing innovations to address the large treatment gap of depression and perinatal depression in LMICs, there is a need for outcome measures that can be used by assessors and health workers not specifically trained in mental health (non-clinicians) in a more structured format. However there is a dearth of locally validated tools for detecting perinatal mental health in Africa.⁶

The AFFIRM-SA randomised controlled trial (RCT) was one such innovation that aimed to address the mental health treatment gap by assessing the effectiveness of a task sharing intervention for a population of women screened positive for perinatal depression in Khayelitsha, a peri-urban informal settlement in Cape Town, South Africa.⁷ The study took place in two primary health care Midwife Obstetric Units (MOUs) in Khayelitsha, an area where levels of crime and unemployment are high,⁸ mental health literacy is low,⁹ and health care at primary level is highly constrained, with few professionals working in highly demanding environments.

The primary outcome measure for this trial was the Hamilton Depression Rating Scale (HDRS).¹⁰ Since it was developed in the 1960s, the HDRS has been acknowledged as a gold

standard for the assessment of depression¹¹ and because of its wide use, is useful to measure change across time periods and countries. It has been used in both the original and in modified structured versions. A meta-analysis of 409 studies across a variety of populations that included coefficients of reliability of the HDRS reported good overall levels of internal consistency (0.784, 0.778–0.789, 95%CI) inter-rater intra-class coefficient (ICC) (0.942, 0.938–0.947, 95%CI), and test–retest reliability ICC (0.93, 0.88–0.96, 95%CI).¹² The tool has also come under criticism, particularly for its failure to include all symptoms of depression (or at least a more modern operationalisation of depression), the inclusion of symptoms other than depressive symptoms (e.g. anxiety), and uneven weightings given to different symptoms.^{11, 13} The tool was also originally designed for use by trained clinicians using observations and probing questions, speaking English, in high-income countries.

Despite these limitations, the HDRS was selected as the primary outcome measure for the AFFIRM-SA RCT because of its wide use, robust psychometric properties and global comparability. Subsequently, the tool needed to be adapted for use by non-clinical fieldworkers, as they were the personnel who would be conducting the assessments for the trial. It also needed to be translated into isiXhosa, the predominant language spoken in Khayelitsha, and validated for this population of pregnant women and mothers of young babies.

Prince¹⁴ provides guidelines of the measures to be taken in order to validly adapt tools. These include investigation of the cultural relevance of the construct, careful translation and adaptation of the measure, and cognitive interviews and pre-testing of the tool. Furthermore, Kagee et al.¹⁵ recommend testing the reliability and validity of adapted instruments before using them in studies. These guidelines were followed, through several phases, which involved: 1) Review of previous structured versions and initial adaptation, 2) Translation and training, 3) Cognitive testing, and 4) Test-retest and inter-rater reliability assessments. These phases of adaptation and validation of the HDRS are presented below.

2. METHODS

2.1 Phase 1: Review of previous structured versions and initial adaptation

Because the HDRS traditionally relies on clinical judgement, the first step of the adaption involved an extensive literature search to find versions that had been structured for administration by non-clinicians. The review by Williams¹⁶ and meta-analysis by Trajkovi et al.¹² were helpful in identifying the past structured adaptations. The most relevant versions found were: the SIGH-D (Structured Interview Guide for the Hamilton Depression Scale)¹⁷, the SI-HDRS (Structured interview version of Hamilton Depression Rating scale)¹⁸, the GRID-HAMD¹⁹, and the DIS-HRSD²⁰.

None of these versions provided structured question *and* response options for all questions, where raters could read the question and elicit a response from participants; while remaining time-efficient and minimising the need for clinical judgement. We thus reviewed the above mentioned versions and decided to combine the detailed questions from the SI-HDRS¹⁸ with the detailed response options from the SIGH-D.¹⁷ The response options were then

standardised wherever possible, to further reduce the need for a clinical judgement. The scoring and meaning of the original 17 item HDRS were not changed.

The SI-HDRS questions diverge from the original HDRS on two items. The first is that the original HDRS asks only about weight loss, and the SI-HDRS asks about weight loss and gain. The second is that the HDRS asks about symptoms in the past week, rather than the past month, which is the case in the SI-HDRS. For both items we decided to align with the SI-HDRS, in recognition that weight gain is correlated with depression, and that symptoms of depression may fluctuate and may not be present in a single week, but may well be present in the past month.

After adaptation, four out of 17 questions still required judgement from the rater. These were Hypochondriasis and Insight, where the question is read out but response options are not given to the participants; and Agitation and Retardation, which are purely observational in nature. In all four questions, detailed categorical response options were then provided for the rater.

2.2 Phase 2: Translation and training

Personnel—Four field workers were recruited as raters for the validation study, and subsequently retained as the assessors for the larger RCT. The raters were all from the same area (Khayelitsha) and spoke the same language (isiXhosa) as the participants and were also fluent in English. They had all worked in data collection previously but had no mental health training. Their average age was 40, and education levels were high school completion and short course diplomas in data collection and good clinical practice.

Procedure—The combined and adapted tool from Phase 1 was translated by two of the isiXhosa speaking raters over a period of two weeks. This involved a thorough explanation in English of each question and answer by a registered counselling clinician (TD), with a discussion regarding how best to translate the sentences. The translation was then back translated into English by a third isiXhosa speaking rater, to ensure reliability of the translation.

Following translation, the raters were trained for a week to administer the tool. This was a critical process in allowing for further clarifications in the translations, following the raters' deeper understanding of the concepts. The training of the raters further confirmed the need for very specific and explanatory response options, beyond the simple 'never' to 'incapacitating' responses in the SI-HDRS and original HDRS. See Appendix 1 for the adapted tool.

2.3 Phase 3: Cognitive testing

The third step in the adaptation process consisted of assessing construct and content validity of the instrument, through cognitive testing.²¹ Cognitive testing aims to evaluate the congruency between the intended meaning of the questions and the interpretation and subsequent response provided by the respondents. The Agitation and Retardation items, which rely on rater judgement and not participant response, were not included in the testing.

Setting and Sample—Cognitive testing was conducted by two raters, through semi-structured interviews with 16 adult pregnant women during one of their antenatal visits at one MOU. Purposive sampling was used to recruit women attending the MOU for their antenatal visits, who were 18 years or older (mean age 25), at any stage in their pregnancy, and able to give informed consent. Both depressed and non-depressed women (EPDS mean: 12.1, AFFIRM-HDRS mean: 12.4) were interviewed over the course of 10 days.

Instruments and procedure—The procedure involved administering the AFFIRM-HDRS to the participants and recording their answers on the score sheet. Once this was completed, the raters read the participants responses back to them one at a time, and they were asked the following probing questions: 1. “Can you explain why you answered in that way?” 2. “What do you think the question is asking about?” and 3. “Can you give me some examples of what you mean?”

Analysis—All of the participants’ answers were audio recorded and later transcribed into English. Answers were analysed by examining the congruency between narratives provided by the respondents to explain the response given, and whether the understanding was ‘in scope’ relative to the original question intent. An item performed well in the analysis when there were none or few repetitions required, most responses were congruent with narratives, and the majority of narratives were in scope with the question.

2.4 Phase 4: Inter-rater and test-retest assessments

Following adaptation, two samples were then recruited to assess the inter-rater, test-retest and internal reliability, and concurrent validity of the AFFIRM-HDRS. The same four raters conducted the assessments, and using feedback from the cognitive testing, another day of training was provided. The registered counsellor (TD) also sat in on initial interviews to provide support if needed.

Setting and Sample—The samples were recruited on two occasions and consisted of isiXhosa-speaking women attending antenatal and postnatal check-ups at the two MOUs identified for the AFFIRM RCT. An average of 30 women attend these clinics per day. Women who were at any stage in their pregnancy or had babies of up to one year old, were eligible. All women had to be 18 years or older, speak isiXhosa, live in Khayelitsha and be able to provide written informed consent. All eligible women were approached, and sampling depended solely on the above criteria. Sampling was deemed representative of the required study population and was conducted until the required sample size was reached.

Sample 1 was recruited for inter-rater reliability over a period of six months, at both MOUs. The sample consisted of 131 pregnant women, which was above the recommended 100, based on an expected relative error margin of 20%, assuming raters will agree 50% of the time ²². **Sample 2** was recruited for test-retest reliability over a period of eight months, after the inter-rater sample, at both MOUs, and consisted of 140 participants. Sample size was calculated using 95% power and an expected reliability of 0.75 ²². Samples 1 and 2 were combined to assess concurrent validity and internal reliability.

Instruments—All women signed informed consent forms and were administered basic demographic questions, the EPDS²³ and the AFFIRM-HDRS. The EPDS screens for symptoms of perinatal depression, is used in clinical and research settings globally, and has been validated in South Africa²⁴, and for antenatal women²⁵. Although it is not a diagnostic instrument, it is a recommended tool for the detection of perinatal depressive symptoms, and it was thus administered in conjunction with the AFFIRM-HDRS, to assess concurrent validity. Women who scored 13 or more on the EPDS, the recommended cut-off for risk of depression²³, were systematically referred to organisations in the community who offer counselling and support for mental health and pregnancy, or the mental health nurse in the clinic.

Procedure

a) Inter-rater reliability: Inter-rater reliability was assessed using the ‘joint interview method’.²⁶ Three pairs of raters were tested, comprising four raters, who interviewed 131 pregnant women. Fifty women were rated by Pair 1, 40 by pair 2, and 41 by pair 3. The process involved both raters being present during the interview and rating the participant at the same time. The primary rater conducted the interview, while the second rater observed and rated the AFFIRM-HDRS separately, but remained silent. The two raters took the role of primary rater an equal number of times, in order to collect data from each rater. Observations for each pair of raters were combined into one dataset, and analyses compared all primary raters (rater 1) against all secondary raters (rater 2).

b) Test-retest reliability: Data collection for test-retest reliability involved conducting the same test to the same participant with the same interviewer within 7 days but no less than one day apart. This was deemed an appropriate length of time given the changing nature of depression, and the difficulty of accessing participants at primary care clinics. The average number of days between assessments was 5.02 (SD: 1.90). The second assessment included a question asking whether there had been any major changes or disturbances in the participant’s life since the last assessment. If there had been changes (e.g. a birth, relationship problems, or death), the participant was excluded from the analysis.

Analysis—Data were entered on Excel, and then exported to STATA 13 for analysis.²⁷ Concurrent validity of the tool was measured by comparing the relationship between the AFFIRM-HDRS scores and EPDS scores across the different samples. Because the AFFIRM-HDRS and EPDS scores were not normally distributed, we conducted a nonparametric test, using Spearman’s Rho. Internal consistency was examined using Cronbach’s Alpha. Inter-rater reliability of the total AFFIRM-HDRS score was assessed for each pair in Sample 1 with the intra-class coefficient (ICC). Cohen’s Kappa was used to assess the overall percentage agreement for each individual item of the AFFIRM-HDRS. This was not calculated for pair’s items where one rating predominated by 90% or more, as the little variance for these items would not result in a meaningful measure of inter-rater reliability.²⁸ Test-retest reliability in Sample 2 for individual items and overall score was also assessed with the intra-class coefficient.

3. RESULTS

3.1 Phase 2

During translation, it was found that in some cases there weren't equivalent isiXhosa words for the clinical terms used in the HDRS. For these words, the language was modified to make it easier for both raters and participants to understand. For instance, "hypochondriacal delusions" (Q.12) was changed to "having false beliefs about your body", and "incapacitating" (Q11), to: "you can't do anything because of it".

Similarly, we attempted to include local idioms for depression in isiXhosa identified in the formative study for the AFFIRM RCT. These included isiXhosa phrases for: stress, thinking too much, being scared, and being sad or unhappy.²⁹ However, the only phrase that could be included in the translation was 'being sad or unhappy'. The rest of these did not fit accurately with the items in the HDRS and were not included, to ensure that the original meaning of the tool did not get lost in the modifications.

Another example of the changes made during translation is that of the question on Guilt (Q9). The most severe response option on the SIGH-D was: "Hears accusatory or denunciatory voices and/or experiences threatening visual hallucinations", and on the SI-HDRS was: "Is experiencing accusatory hallucinations, i.e., voices and/or delusions". The language and phrasing of the AFFIRM-HDRS was adapted instead to: During the past month, have you been feeling guilty or bad about something you have done? Do you feel you have let people down or that you are evil? Do you think your illness is punishment for something? Response options: (0) No feelings of guilt; (1) You feel you have let people down OR you feel evil or bad; (2) You feel you have let people down AND you feel evil or bad; (3) You think that your illness is a punishment; (4) You feel like you are hearing voices in your head or you feel you are so bad that you will hurt other people, or it will lead to your own death.

3.2 Phase 3

In the cognitive analysis, eight out of the 15 tested items performed well. The remaining seven required some explanation by raters or were misinterpreted by a few respondents. Two items required wording changes. The first was Q1, where participants tended to interpret the word 'work' as paid employment, rather than to include usual home activities too. This population is largely unemployed and thus house work and child care was their primary 'work', but they answered that they had no decrease in 'work' activities because their depression had not affected their (non-existent) employment. The question was therefore adapted to include "your activities *or house work*".

The second was the understanding of the isiXhosa word for "fatigue" in Q8, which was changed to "tiredness or loss of strength". The word "aches" was also changed to "body aches", and then this was translated into isiXhosa as "headache, backache, or pains in certain parts of your body."

There was also confusion for some participants around the different time frames of the three insomnia questions. To explain this, an introductory statement was added to clarify the difference between the three stages of sleep.

Lastly It was found that there was confusion in the options on the Insight responses: between acknowledging the illness or denying the illness. Retraining for this item involved suggesting that the raters kept in mind whether the participant was “telling the truth” about her illness, which helped their comprehension of the question. The response option in isiXhosa was changed to: “She says she is not depressed, but she is depressed at the moment in my opinion”.

3.3 Phase 4

Sample characteristics—The demographic characteristics of participants recruited in Sample 1 and 2 are presented in Table 1. Participants in both samples were similar; they had an average age of 27 years, the majority reported having partners, and nearly half were employed. Mean EPDS scores were higher in Sample 1 (mean=10.4, SD=5.20) compared to Sample 2 (mean 5.3, SD=4.71). It is unclear why depression scores are lower in Sample 2, than Sample 1, other than that they were collected at different time points.

Concurrent validity—For Sample 1, the mean AFFIRM-HDRS scores were significantly associated with EPDS scores ($Rho=0.60$, $p<.001$). Sample 2 had slightly lower associations with a Rho of 0.43 , $p<.001$.

Reliability

a) Internal consistency: The overall internal consistency for the AFFIRM-HDRS in Sample 1 was an alpha of .77, and .71 in Sample 2. The combined samples had a Cronbach’s alpha of .74. Item-rest correlations (correlations of each item with the whole scale with that item excluded) for individual items in Sample 1 were good for all but five items scoring above .30. The exceptions were: Weight loss or gain, Libido, Insight, Retardation and Agitation. Exceptions for Sample 2 were: Work and activities, Weight loss or gain, Appetite, Libido, and Insight. Those lowest across both samples were: Weight loss or gain, Libido, and Insight (Table 2).

b) Inter-rater reliability: The intra-class coefficients (ICC) for the AFFIRM-HDRS were very similar and almost perfect for the three pairs of raters, ranging from .97 (.94 – .99) for pair 3, to .98 (.97 – .99) for pair 1 and 2. The overall percentage agreement and Kappa figures for individual items of the AFFIRM-HDRS for all three pairs are presented in Table 3. Because of too little variation in ratings, Kappa coefficients were not calculated for item 17 for all three pairs, for item 12, 13 and 14 for Pair 2, and for items 13, 14 and 16 for Pair 3. A Kappa coefficient for item 16 was also not calculated for Pair 2, as both raters rated 0 for all participants.

For the individual items, all but two Kappa coefficients were deemed almost perfect, ranging from .86 to 1.00³⁰. Only Insight (for pair 1 and 3), and Retardation (for pair 1) showed a lower, but still acceptable agreement (of .64, .79 and .44 respectively). All other items

showed overall percentage agreement ranging from .94 to 1.00, across pairs. Altogether, raters only differed on their overall ratings of the AFFIRM-HDRS between 3.24% and 6.24% of the time.

c) Test-retest reliability: Test-retest reliability ICC figures are presented in Table 4. Results showed excellent overall ICC reliability over time (.90; 95% CI: .86-.93). Item level retest reliability had more variance, and ranged from .42 - .90. The most problematic item was Weight loss or gain, at .42. The other four moderate items were Appetite, Somatic anxiety, Insight and Agitation (.63 - .66). The items with the highest retest reliability were General somatic symptoms (.90), Early insomnia (.87), and Late insomnia (.86).

4. DISCUSSION

A straightforward translation of an existing scale is inadequate for use in different linguistic and cultural situations, and careful adaptation, translation and validation of an existing measure is necessary to be culturally sensitive but also globally comparable. This study set out to adapt and validate the structured AFFIRM-HDRS to be used by non-clinical raters in a low-income setting, in isiXhosa, for the AFFIRM-SA RCT. First, it was decided that previous versions of the tool, although structured, were not sufficient for non-clinicians to use, and thus we combined these versions and created a more explanatory and structured version. Furthermore, the items in the HDRS do not lend themselves to simple translation into isiXhosa, or necessarily to understanding by participants or raters. In fact, transforming a clinical interview with clinician-directed questions into a form that can be understood and *explained* by non-clinicians takes careful and considerate training.

Cognitive testing is also a very important process in the adaptation of a screening tool.³¹ Participants themselves are very rarely asked how they interpret the questions they are asked and what meaning they give them, and this is particularly important for a tool developed in a western and English context. Similarly, the cognitive testing also highlighted the few misinterpretations that were being made by the raters themselves, which required clarification and retraining. The cognitive testing established congruency of intended meaning of the questions and interpreted meaning by participants and demonstrated adequate construct and content validity of the adaptation.

Phase four of the research assessed the measure of agreement and consistency of the tool across raters and over time. The tool had significant associations with the widely used EPDS, which demonstrated acceptable concurrent validity. Overall, reliability, consistency over time, and reproducibility of the AFFIRM-HDRS, was found to be highly acceptable.

When examining how well items were intercorrelated, the tests showed good overall internal consistency,³² with Cronbach's Alpha ranging from .71 to .77 across subsamples. However, individual item-rest correlations suggested that certain items did not fit well with the overall instrument, particularly Weight loss or gain, Insight, and Libido. The meta-analysis by Trajković et al.¹² also reported that Insight and Libido had 'barely adequate' item total correlations.

Regarding the Libido item, there may have been a culturally induced modesty in answering this question, which may have lead to under-reporting of symptoms. However, the cognitive testing demonstrated that the item performed very well, with no misinterpretations or incongruent responses. It is therefore unclear why the item performed so badly in the reliability and validity testing and should be examined in further research.

Another possible reason for the above 'lack of fit' of the three items is that the HDRS is a multidimensional instrument,^{12, 33} and measures a variety of dimensions of depression rather than being a unidimensional measure. A factor analysis may have been able to group items from the AFFIRM-HDRS into various dimensions to identify similar variables, but this was beyond the scope of the current paper. However, other studies have conducted factor analyses in the past, with similar findings to each other. The most common items identified for a more unidimensional measure of major depression are reported as: Depressed mood, Guilt, Suicide, Psychic anxiety, Work and activities, Agitation and Libido.^{34–36} In a similar vein, Bech et al.³⁷ conducted a Rasch analysis of each item of the HDRS, and proposed that a six item scale may be more accurate in measuring depression severity than the original HDRS. These six items were: Depressed mood, Guilt, Psychic anxiety, Work and activities, Retardation, and General somatic symptoms.³⁷

Tests of reliability between raters showed extremely high kappa coefficients (ranging from .86 to 1.00), demonstrating good representations of the variables in the adapted tool. The only exceptions here were Insight (.64) and Retardation (.44). In the AFFIRM-HDRS, the Insight and Retardation items require the highest degree of 'clinical' judgement from the rater, which may be one reason for their poorer reliability across raters and tests. In addition, the Insight item may have been misunderstood because of the lack of mental health literacy and recognition of symptoms of depression in this population. This item had also already been modified after the cognitive testing, and thus could not be further altered for risk that the tool as a whole may not then be comparable with other interventions that have used the HDRS. In other studies examining the HDRS, Insight is also consistently the least reliable item.^{11, 12} This demonstrates that this item may be a limitation of the original HDRS across versions, and not specifically of the AFFIRM-HDRS.

In their review of the HDRS in 2004, Bagby et al. write that it is necessary to demonstrate that scores remain the same over time in the absence of treatment, if they are expected to change in response to effective treatment. The test-retest results for the AFFIRM-HDRS show the overall intraclass coefficient to be very good, at 0.90,^{30, 38} which is a heartening sign of consistency of the overall tool. It also falls within the range that both Bagby et al.¹¹ and Trajkovi et al.¹² reported in their reviews of multiple studies; with the former having overall retest reliability from 0.81 to 0.98, and the latter a mean figure ranging between 0.65 and 0.98.

Item level reliability for the AFFIRM-HDRS test-retest results had a wider range of scores, with the lowest scoring item being Weight loss or gain, at .42. All other items scored between .63 and .90. This is higher than international comparisons, which range from .00 to .85.¹¹ The Weight item also had low retest reliability in the SI-HDRS. The poor test-retest reliability (and item-rest correlation) of the 'Weight loss or gain' item in the AFFIRM-

HDRS may indicate that the question does not discriminate well between weight gain related to depression or to pregnancy. However, the item performed well on the cognitive testing, which implies that the participants themselves understood the question. Thus possibly 'gaining weight' should not have been included in the option and the question should be reverted to the original HDRS which only asks about 'weight loss'.

With a mean internal consistency alpha of .74, mean intraclass correlations across raters of .98, and over time, of .90, the AFFIRM-HDRS performs very similarly to the GRID-HAMD, and to a large degree has better reliability and reproducibility than the SIGH-D and SI-HDRS versions. The SIGH-D had a mean retest reliability of .81, but with 13 items not meeting the criteria for adequate reliability. Inter-rater reliability was not reported. THE SI-HDRS had high inter-rater ICC of .92, but this was conducted by psychiatrists, and only on 20 participants. Their ICC over time involved nurses, and this was lower, at .65, with only one item scoring above .70 (Early insomnia). This may demonstrate the benefits and effectiveness of the more structured, and therefore reproducible nature of the AFFIRM-HDRS adaptation, which, as argued by Williams,¹⁷ would reduce room for variance in interpretation, over time.

5. LIMITATIONS

As almost all of the participants were pregnant at the time of interviewing, they may have used this as their frame of reference in thinking about and answering the questions, particularly those referring to physical symptoms. The cognitive testing identified the four symptoms most often described with reference to pregnancy. These were: Weight loss or gain, Guilt, Somatic anxiety (physical symptoms) and Hypochondriasis. These responses were flagged as issues to be aware of in the broader RCT primary outcome analysis.

There was also too little variation in the inter-rater responses to calculate Kappa coefficients for five items across the three pairs. Monitoring of scores and subsequent purposive sampling to recruit women with a wider range of severity scores could have helped with generating data with greater variability on these items.

As much as the tool attempted to be locally valid, there is always room for cultural or linguistic misinterpretation, which is difficult to assess or identify, beyond the process of cognitive testing. However, we believe that by following guidelines provided by Prince¹⁴ and Kagee et al.,¹⁵ and being committed to a thorough and extensive translation process, we have minimised the risk for linguistic or cultural misinterpretation to a large degree.

Although some items had lower item-rest correlations with the tools as a whole, we believe the tool is relevant in assessing depressive symptoms amongst perinatal women within Khayelitsha, because of the extensive translation process and constant checking with the non-clinicians, and the cognitive testing with the participants.

Lastly, it is important to remember that the aim of this study was to adapt the diagnostic tool to measure local clinical severity in this particular population through a continuous measure, and not diagnose depression itself. It is hoped that this tool will assist with adequate referrals to existing systems for those with more severe symptoms. Future research would be useful to

validate the AFFIRM-HDRS against a gold standard diagnostic instrument, to determine criterion validity.

6. RECOMMENDATIONS

We are satisfied that it is possible for non-clinicians to conduct this adapted AFFIRM-HDRS, on condition that adequate training and supervision is provided. For use in different contexts, we recommend the comprehensive process we followed, to adapt and validate an international tool to be relevant and contextual to the situation it is used in, as described in the methods section. We also found that the process of back and forward translation by the non-clinicians was a good training tool in itself, and also provided as a means to assess their levels of understanding of the tool. We would thus recommend this process in training of raters in future studies. Lastly, we recommend removing 'weight gain' from the 'weight loss or gain' item in particular for assessing pregnant women.

The English version of the AFFIRM-HDRS is available in Appendix 1, and the isiXhosa version is available on request.

7. CONCLUSION

We believe that the AFFIRM-HDRS is a robust, reliable and valid tool for use by non-clinicians for assessing depression in perinatal isiXhosa populations. It is hoped that the adaptation and validation of this internationally recognised measurement tool can assist with a more accurate detection and assessment of depression in this local setting, and will play a role in improving mental health services and addressing the treatment gap for depression. It could also be useful for measuring changes in depression symptom severity in the increasing number of intervention studies in mental health in low and middle income countries, and it could enable workers not specialised in mental health, such as nurses or lay-counsellors to identify depression in resource constrained environments where often no assessment is done at all. The AFFIRM-HDRS thus fills the need for a measurement tool for depression that is entirely structured so that it can be satisfactorily conducted by non-clinicians.

Acknowledgements

This research could not have been done without the devotion and dedication of Vuyiseka Shumana, Phatiswa Mhlebi, Siphokazi Nyebilele and Phumza Gqwaka. Thank you.

Funding

This research was conducted as part of the Africa Focus on Intervention Research for Mental Health (AFFIRM). Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health (NIH) under Award Number U19MH095699. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix 1

AFFIRM-HDRS:

Structured Hamilton Depression Rating Scale

Instructions: Score answers by the number in front of the question.

These questions relate to your feelings over the past month.

1 WORK AND ACTIVITIES

During the past month, have you been less able than usual to work or do your usual activities? Do your activities make you feel tired, or have you lost interest in your activities or house work?

- 0 - No decrease in productivity or time spent at work and/or doing usual activities
- 1 - Activities make you feel tired
- 2 - Lost interest in work or activities or house work
- 3 - Decrease in productivity of work or activities or house work
- 4 - Spending less time at work or doing activities

2 WEIGHT LOSS

Have you lost any weight during the past month?

- 0 - No
- 1 - Possibly
- 2 - Yes, definite change in weight (not on diet)

3 APPETITE

During the past month, have you experienced a loss of appetite?

- 0 - No loss of appetite
- 1 - Some loss of appetite but you are still eating
- 2 - At least some loss of interest in food and you require encouragement to eat

4 LIBIDO

During the past month, have you had an interest in sex?

- 0 - Yes, normal interest in sex (or is not sexually active)
- 1 - Somewhat less interest
- 2 - A lot less interest than usual or no interest at all

INSOMNIA:

The following 3 questions will ask you about different aspects of sleep: going to sleep at night, waking up in the middle of the night, and waking up early or sleeping late in the morning.

5 EARLY INSOMNIA

Firstly, during the past month, have you had troubles or difficulties falling asleep?

- 0 -** No
- 1 -** Sometimes
- 2 -** Yes, almost every night

6 MIDDLE INSMONIA

During the past month, have you been waking up during the night?

- 0 -** No
- 1 -** Sometimes
- 2 -** Yes, almost every night and you get out of bed (other than for urinating/peeing)

7 LATE INSOMNIA

During the past month, have you either been waking up earlier in the morning than you wanted to, or sleeping too much?

- 0 -** No
- 1 -** Sometimes
- 2 -** Yes, most of the time

8 GENERAL SOMATIC SYMPTOMS

During the past month, have you experienced fatigue or had less energy than usual? Or have you had headaches, backaches, or aches in specific parts of your body?

- 0 -** None
- 1 -** Some
- 2 -** Yes, a lot

9 GUILT

During the past month, have you been feeling guilty or bad about something you have done? Do you feel you have let people down or that you are evil? Do you think your illness is punishment for something?

- 0 -** No feelings of guilt
- 1 -** You feel you have let people down OR you feel evil or bad
- 2 -** You feel you have let people down AND you feel evil or bad
- 3 -** You think that your illness is a punishment
- 4 -** You feel like you are hearing voices in your head, or you feel you are so bad that you will hurt other people, or it will lead to your own death

10 PSYCHIC ANXIETY

During the past month, have you been feeling nervous, anxious, worried or frightened?

- 0 -** Never

- 1 - Sometimes
- 2 - Quite often
- 3 - Most of the time
- 4 - Yes, signs or problems are very bad all of the time

11 SOMATIC ANXIETY

Now I am going to read you a list of physical symptoms. Tell me if you have experienced any of these in the last month and how severe they are:

- A. Stomach or digestive problems or pains
- B. Heart palpitations
- C. Breathing very fast or trouble breathing
- D. Urinating often
- E. Muscle aches, body aches
- F. Unusual sensations like trembling or ringing in your ears
- G. Flushing, feeling faint, or sweating

Code for the most severe symptom:

- 0 - Absent
- 1 - A little bit
- 2 - Some
- 3 - A lot
- 4 - Severe, and you can't do anything because of it

12 HYPOCHONDRIASIS

During the past month, have you been worrying more than usual about your health and how your body is working? (Apart from normal fears about your pregnancy). [Don't read the options to the participant. Make the assessment yourself.]

- 0 - No unnecessary worries about her health
- 1 - Some unnecessary worry about her health
- 2 - A lot of unnecessary worries about her health
- 3 - Strong beliefs she has a physical problem and doctors won't believe her
- 4 - Delusional, e.g. she thinks that her body is rotting

13 SUICIDE

During the past month, have you had thoughts that life is not worth living, or that you would rather be dead? Have you had thoughts of hurting or killing yourself? If yes, how often?

- 0 - No

- 1 - Sometimes
- 2 - Often
- 3 - Most of the time
- 4 - Suicide attempt

14 INSIGHT

Do you think that you have a psychological problem, such as depression? [Don't read the options to the participant. Make the assessment yourself.]

- 0 - Acknowledges being depressed or having a psychological problem (OR is not currently depressed)
- 1 - Acknowledges illness but blames it on something else
- 2 - Denies any illness but is currently depressed in your opinion

15 DEPRESSED MOOD

During the past month, have you been feeling sad, depressed, helpless, hopeless, or worthless? If yes, how often do you feel this way?

- 0 - No, not at all
- 1 - Occasionally
- 2 - Quite often
- 3 - Very often
- 4 - Yes, almost all the time

16 RETARDATION

Observe and rate slowness of thought, speech, concentration, and physical movement [Don't read the options to the participant. Make the assessment yourself.]

- 0 - Normal speech and thought
- 1 - A bit of slowness in thinking or speaking (Slight retardation)
- 2 - A lot of slowness in thinking or speaking (Obvious retardation)
- 3 - A lot of very long pauses (Interview difficult)
- 4 - Interview impossible

17 AGITATION

Observe and rate restlessness, fidgetiness and physical activity [Don't read the options to the participant. Make the assessment yourself.]

- 0 - None
- 1 - Fidgetiness
- 2 - Playing with hands, hair, obvious restlessness

- 3 - Restless, moving about; can't sit still
- 4 - Hand wringing, nail biting, hair pulling, biting of lips, person is moving about a lot

TOTAL SCORE _____

8. REFERENCES

1. Whiteford H, et al., Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet*. 2013 382: p. 1575–1586.
2. Fisher J, et al., Prevalence and determinants of common perinatal mental disorders in women in low- and lower-middle-income countries: a systematic review. *Bull World Health Organ*. 2012 90(2): p. 139G–149G.
3. Kohn R, et al., The treatment gap in mental health care. *Bulletin of the World health Organization*. 2004 82(11): p. 858–866. [PubMed: 15640922]
4. Patel V, de Souza N, and Rodrigues M, Postnatal depression and infant growth and development in low income countries: a cohort study from Goa, India. *Arch Dis Child*. 2003 88: p. 34–37. [PubMed: 12495957]
5. Rahman A, Harrington R, and Bunn J, Can maternal depression increase infant risk of illness and growth impairment in developing countries? *Child Care Health Dev*. 2002 28(1): p. 51–56. [PubMed: 11856187]
6. Tsai AC, et al., Reliability and validity of instruments for assessing perinatal depression in African settings: systematic review and meta-analysis. *PLoS One*. 2013 8(12): p. e82521.
7. Lund C, et al., Task sharing of a psychological intervention for maternal depression in Khayelitsha, South Africa: study protocol for a randomized controlled trial. *Trials*. 2014 15: p. 457. [PubMed: 25416557]
8. StatisticsSA, Statistics in Brief, S.S. Africa, Editor. 2014: Pretoria.
9. Dirwayi N, Mental illness in primary health care: a study to investigate nurse's knowledge of mental illness and attitudes of nurses toward the mentally ill., in *Psychology Masters Thesis*. 2002 University of Cape Town.
10. Hamilton M, A rating scale for depression. *J. Neurol. Neurosurg. Psychiatr* 1960 23(56).
11. Bagby M, et al., The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight? *Am J Psychiatry*. 2004 161.
12. Trajkovi G, et al., Reliability of the Hamilton Rating Scale for Depression: a meta-analysis over a period of 49 years. *Psychiatry Res*. 2011 189(1): p. 1–9. [PubMed: 21276619]
13. Cusin C, et al., Rating scales for depression, in *Handbook of clinical rating scales and assessment in psychiatry and mental health*. 2009, Springer p. 7–35.
14. Prince M, Measurement validity in cross-cultural comparative research. *Epidemiologia e Psichiatria Sociale*. 2008 17(03): p. 211–220. [PubMed: 18924560]
15. Kagee A, et al., Screening for common mental disorders in low resource settings: reasons for caution and a way forward. *Int Health*. 2013 5(1): p. 11–4. [PubMed: 23580905]
16. Williams JB, Standardizing the Hamilton Depression Rating Scale: past, present, and future. *European Archives of Psychiatry and Clinical Neuroscience*. 2001 251: p. 6–12. [PubMed: 11315519]
17. Williams JW, A structured interview guide for the hamilton depression rating scale. *Archives of General Psychiatry*. 1988 45(8): p. 742–747. [PubMed: 3395203]
18. Potts M, et al., A structured interview version of the Hamilton depression rating scale: Evidence of reliability and versatility of administration. *J Psychiat Res*. 1990 24(4): p. 335–350. [PubMed: 2090831]
19. Williams J, et al., The GRID-HAMD: standardization of the Hamilton Depression Rating Scale. *International Clinical Psychopharmacology*. 2008 23: p. 120–129. [PubMed: 18408526]

20. Whisman M, et al., A Structured Interview Version of the Hamilton Rating Scale for Depression: Reliability and Validity. *Psychological Assessment*. 1989 1(3): p. 238–241.
21. Willis GB, Cognitive interviewing and questionnaire design: a training manual. 1994: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.
22. Gwet KL, Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters. 2012: Advanced Analytics Press.
23. Cox JL, Holden JM, and Sagovsky R, Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. *The British Journal of Psychiatry*. 1987 150(6): p. 782–786. [PubMed: 3651732]
24. de Bruin G, et al., The factor structure of the Edinburgh Postnatal Depression scale in a South African peri-urban settlement. *S Afr J. Psychol* 2004 34(1): p. 113–121.
25. Cox J. and Holden J, Perinatal Psychiatry: Use and Misuse of the Edinburgh Postnatal Depression Scale: Papers. 1994: Gaskell.
26. Bruss G, et al., Hamilton anxiety rating scale interview guide: Joint interview and test-retest methods for interrater reliability. *Psychiatry Research*. 1994 53(2): p. 191–202. [PubMed: 7824679]
27. StataCorp, Stata Statistical Software: Release 13. 2013 StataCorp LP: College Station, TX.
28. Feinstein AR and Cicchetti DV, High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology*. 1990 43(6): p. 543–549. [PubMed: 2348207]
29. Davies T, et al., “The sun has set even though it is morning”: Experiences and explanations of perinatal depression in an urban township, Cape Town. *Transcultural Psychiatry*. 2016 53(3): p. 6.
30. Landis JR and Koch GG, The measurement of observer agreement for categorical data. *biometrics*. 1977: p. 159–174. [PubMed: 843571]
31. Mehrotra S, Cognitive interviewing: An overview and an illustration. *Journal of the Indian Academy of Applied Psychology*. 2007 33(1): p. 81–84.
32. Nunnally J. and Bernstein I, *Psychometric Theory* McGraw-Hill New York Google Scholar. 1978.
33. Pancheri P, et al., Psychopathological dimensions of depression: a factor study of the 17-item Hamilton depression rating scale in unipolar depressed outpatients. *Journal of Affective Disorders*. 2002 68(1): p. 41–47. [PubMed: 11869781]
34. Gibbons RD, Clark DC, and Kupfer DJ, Exactly what does the Hamilton depression rating scale measure? *Journal of Psychiatric Research*. 1993 27(3): p. 259–273. [PubMed: 8295158]
35. Faries D, et al., The responsiveness of the Hamilton Depression Rating Scale. *Journal of Psychiatric Research*. 2000 34(1): p. 3–10. [PubMed: 10696827]
36. Maier W, Philipp M, and Gerken A, Dimensions of the Hamilton depression scale. Factor analysis studies. *European archives of psychiatry and neurological sciences*. 1985 234(6): p. 417–422. [PubMed: 4029226]
37. Bech P, et al., Quantitative rating of depressive states: Correlation Between Clinical assessment, Beck’s Self-Rating Scale and Hamilton’s Objective Rating Scale. *Acta Psychiatrica Scandinavica*. 1975 51(3): p. 161–170. [PubMed: 1136841]
38. Anastasi A. and Urbina S, eds. *Psychological Testing*. 7 ed. 1997, MacMillan: New York./

Table 1:

Demographic characteristics of Samples 1 and 2.

Demographics	Sample 1 (Inter-rater)		Sample 2 (Test-retest)	
	100%	n = 131	100%	n = 140
Mean (SD) age (years)	27.2 (5.83)		26.7 (5.43)	
Partner status:				
Has partner but lives separately	56.5%	74	40.1%	56
Has partner and lives together	40.5%	53	28.9%	40
Single and lives with their family	3.0%	4	28.9%	40
Single and lives alone	0%	0	2.1%	3
Education:				
Completed final school year	42.8%	56	54.9%	77
Employment:				
Employed (full or part time)	46.6%	61	48.6%	68
Unemployed	53.4%	70	51.4%	72
Pregnant (yes/no):	100%	131	90.9%	127
Mean (SD) EPDS score	10.4 (5.20)		5.3 (4.71)	
Mean (SD) AFFIRM-HDRS score	14.5 (5.97)		11.4 (4.11)	

Table 2:

Item-rest correlation (Cronbach's alpha) across samples

	Item	First Raters	Second Raters	Test 1	Test 2
1	Work and activities	.3557	.3415	.1520	.0821
2	Weight loss or gain	.1861	.1742	.0200	.0443
3	Appetite	.3514	.3238	.1418	.1900
4	Libido	.2041	.2567	.0826	.1160
5	Early insomnia	.5180	.5251	.4389	.4187
6	Middle insomnia	.5738	.5537	.3285	.3849
7	Late insomnia	.3152	.4219	.3053	.2973
8	General somatic	.3747	.3242	.3714	.4245
9	Guilt	.4178	.4505	.5294	.3819
10	Psychic anxiety	.4840	.4871	.5343	.4745
11	Somatic anxiety	.3520	.3588	.4475	.4164
12	Hypochondriasis	.5148	.5158	.4884	.4947
13	Suicide	.4427	.4728	.3738	.4010
14	Insight	.0529	.0534	.1601	.2942
15	Depressed mood	.4551	.4583	.5198	.4834
16	Retardation	.1639	.3391	.4308	.4256
17	Agitation	.3129	.2796	.3822	.3583

Table 3:

Inter-rater agreement for individual items

Item		Pair 1		Pair 2		Pair 3	
		P_o^a	K^b	P_o^a	K^b	P_o^a	K^b
1	Work and activities	.92	.89	.90	.86	1.00	1.00
2	Weight loss or gain	.96	.90	1.00	1.00	.98	.95
3	Appetite	.94	.91	1.00	1.00	1.00	1.00
4	Libido	1.00	1.00	.98	.96	1.00	1.00
5	Early insomnia	.98	.97	.93	.87	1.00	1.00
6	Middle insomnia	.90	.84	.95	.90	.98	.94
7	Late insomnia	.88	.82	.98	.96	1.00	1.00
8	General somatic	.98	.96	.98	.96	.95	.92
9	Guilt	.98	.96	.95	.92	.98	.95
10	Psychic anxiety	.98	.97	.95	.91	.98	.95
11	Somatic anxiety	.92	.88	.95	.93	.93	.88
12	Hypochondriasis	.94	.89	.95	-	.98	.93
13	Suicide	.96	.88	1.00	-	1.00	-
14	Insight	.84	.64	.95	-	.79	-
15	Depressed mood	1.00	1.00	1.00	1.00	.95	.89
16	Retardation	.84	.44	1.00	-	.88	-
17	Agitation	.92	-	1.00	-	.93	-
	Overall AFFIRM-HDRS	.94		.97		.96	

^aOverall percentage agreement^bCohen's Kappa

Table 4:

Test-retest reliability (ICC) for individual items

Item	ICC	95% Conf. Interval	
Work and activities	.8132	.7491	.8622
Weight loss or gain	.4236	.2790	.5494
Appetite	.6461	.5393	.7324
Libido	.8145	.7509	.8632
Early insomnia	.8663	.8186	.9021
Middle insomnia	.8134	.7495	.8624
Late insomnia	.8612	.8119	.8983
General somatic	.9037	.8684	.9298
Guilt	.8111	.7464	.8606
Psychic anxiety	.8133	.7493	.8623
Somatic anxiety	.6310	.5210	.7204
Hypochondriasis	.7904	.7197	.8448
Suicide	.7603	.6813	.8218
Insight	.6650	.5625	.7475
Depressed mood	.8187	.7563	.8663
Retardation	.7940	.7243	.8476
Agitation	.6539	.5489	.7387
Overall AFFIRM-HDRS	.8999	.8633	.9271