*Martin Bernstorff | #201307263  
Supervisor: Anders Foldspang*

*18/10-16*

A comparison of the Edinburgh Postnatal Depression Scale (EPDS) and the Postpartum Depression Screening Scale (PDSS) for peripartum depression screening

# Formalities

Indholdsfortegnelse

Formalia 2

Abstract 4

Problem Statement 5

1. Introduction 6

2. Methods 8

Search string 8

2.1 Selection criteria 8

3. Findings 9

3.1 Introduction and aims 9

3.2 Sample characteristics 9

3.3 Protocols 10

3.4 Findings 10

4. Discussion 11

4.1 Test protocol 11

4.1.1 Construct 11

4.1.1.1 Construct similarity 11

4.1.2 Index test 11

4.1.2.1 Index test comparability 11

4.1.2.3 Chronology 11

4.1.3 Reference standard 11

4.1.3.1 Inter-observer variation 11

4.1.3.2 Diagnostic review bias 11

4.1.3.3 Diagnostic test comparability 11

4.2 Study populations 12

4.2.1.1 Demographic features 12

4.2.1.2 Disease prevalence 12

4.2.1.3 Population size 12

4.3 Flow and timing 13

4.3.1.1 Disease progression bias (time difference between index test and reference test) 13

4.3.1.2 Induction of signal by questionnaire 13

4.4 Considerations in selection of optimal cut-off value 13

4.4.1.1 Consequences of a false-positive 13

4.4.1.2 Consequences of a false-negative 13

Conclusion 14

References 15

# Abstract

Peripartum depression affects 3-6% of all women who give birth and is therefore a highly prevalent and import health issue. Given that depression has negative consequences to child, mother and family and that treatment is available, detection of depression is important. Screening is an effective tool for routine detection of low-grade disease, which in the case of peripartum depression can be done by the simple and non-invasive administration of a questionnaire. Two such questionnaires are compared by examination of two comparative studies, but due to methodological differences and insufficient sample sizes no firm conclusions can be drawn. Further research is needed.

# 1. Introduction

Peripartum depression (PPD) refers to the combination of major depressive disorder (MDD) or minor depressive disorder (mDD), as defined in the Diagnostic and Statistical Manual of Mental Disorders V (DSM-V), with the specifier “with peripartum onset”.

Major depressive disorder is defined by the following diagnostic criteria1:

1. “Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.
   1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g. feels sad, empty, hopeless) or observation made by others (e.g., appears tearful).
   2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).
   3. Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly evert day.
   4. Insomnia or hypersomnia nearly every day.
   5. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
   6. Fatigue or loss of energy nearly every day.
   7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
   8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
   9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.
2. The symptoms cause clinically significant distress or impairment in social, occupational, or other areas of functioning.
3. The episode is not attributable to physiological effects of a substance or to another medical condition.”

Minor depressive disorder (mDD) is not specified in the DSM-V, but was specified in the DSM-IV-TR (pp. 320-350)2 as 2 to 4 of the abovementioned symptoms (A1-A9) during a 2-week period.

Both disorders can be appended the specifier “with peripartum onset” if the onset of the current or most recent episode of major depression occurs during pregnancy or in the 4 weeks following delivery1, p. 186. When given the specifier, this thesis will refer to MDD as major peri-partum depression (MPPD) and mDD as minor peri-partum depression (mPPD).

The incidence of MPPD is 3-6%, p. 1861. Estimates vary wildly, and 3-6% seem to be a lower-bound estimate. One can therefore argue that MPPD is an important health problem, as depression carries negative consequences for mother, child and family3. While the aetiology of depression is not well understood, early treatment is more effective due to prevention of future negative consequences and cost-effective treatment is available4.

For early treatment to be effective, identification of the affected individuals is essential. Given the information above, screening is indicated. For screening to be effective, the appropriate diagnostic device must be used. Two such devices are the Postpartum Depression Screening Scale (PDSS) and the Edinburgh Postpartum Depression Scale (EPDS).

The most well-known scale, the EPDS, was developed in 1987 by Cox, Holden and Sagovsky and is now used worldwide by midwives and doctors alike. It is the de facto standard screening tool for PPD but varies widely in its reported sensitivity and specificity{Zelkowitz:1995ul}{Thompson:1998wd}{Lawrie:1998wt}.

The PDSS was developed by Beck and Gable as a result of more than 10 years of qualitative research and has been extensively validated5. As a result, it should better reflect the psychometric properties of women experiencing peripartum depression and therefore have a larger area under the curve (AUC) for a receiver-operator characteristics (ROC) curve when compared to a reference-standard, and thus be better suited for screening.

The reference standard in the literature is the Structured Clinical Interview for DSM-IV6 (SCID).

This thesis will compare the ROC-AUC of the PDSS and the EPDS as regarding to mPPD and MPPD with the aim of determining the most precise questionnaire.

# 2. Methods

### Search string

("screening”[title] AND (“EPDS” OR “Edinburgh Postnatal Depression Scale”) AND ("Postpartum Depression Screening Scale” OR “PDSS”))

AND

(“comparative study”[publication type] OR “combined”[title] OR “comparison”[title] OR “comparative”[title])

AND

(“sensitivity” OR “specificity”)

NOT

(“review”[publication type])

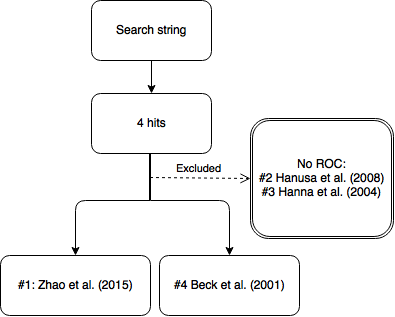
## 2.1 Selection criteria

This is a bachelor’s thesis and must therefore be written within certain boundaries. For this reason, and based on the advice of my advisor, the search-scope has been narrowed extensively.

Searches were performed in the PubMed database.

The search-string consists of 4 blocks.

1. Subject matter.
2. Only comparative studies, as to get comparable results in the same demographic.
3. Studies must use the word sensitivity or specificity in their abstract, to narrow the search scope and increase the chance of the paper featuring a ROC-curve.
4. Reviews were excluded as we were instructed to use only original articles.



This search string returns 4 hits. Articles that did not contain ROC-curves (n = 2) were excluded.

This leaves two suitable articles, Zhao et al. and Beck et al.

# 3. Findings

## 3.1 Introduction and aims

Both Beck7 et al. and Zhao et al.8 share the goal of comparing the

EPDS with the PDSS.

The introduction in Beck et al. bear clear semblance of the psychological background of Gable, presenting a thorough and relevant review of the literature ending in a specific set aims.

Zhao et al. give a more bare-bones review, with a specific focus on China and vaguer aims. As an example, they set out to determine the cut-off scores of the PDSS and the EPDS “if combined use”. This ambiguity is not further expanded upon.

## 3.2 Sample characteristics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Criterion** | **Beck et al. (2001)** | | **Zhao et al. (2015)** | | |
| **Sample size** | 150 | | 842 | | |
| **Inclusion criteria** | · Age ≥ 18  · Able to speak and read English  · 2-12 weeks post-partum  · Delivered a live, healthy infant | | · Obstetric complication  · Pregnant (i.e. antepartum) | | |
| **Recruitment** | · Invitation during “Preparation for child-birth classes” (n = 122)  · Newspaper advertisement (n = 28) | | · Convenience sampling of all women with obstetric complications attending the antenatal clinic at Fudan University Hospital | | |
| **Age (mean ± S.D.)** | 30.5 ± 3.7 | | 31 ± 4.82 | | |
| **College or higher** | 81% | | 87.2% | | |
| **Major ethnicity** | White (87%) | | Asian (100%) | | |
| **Country** | United States | | China | | |
| **Language** | English | | Chinese | | |
|  | **Weeks since delivery (mean ± SD)** | | **Gestational weeks** | | |
|  | 5.6 ± 1.52 | | ≤ 12 | | 6.8% |
| 13-27 | | 49.0% |
| 28-34 | | 39.7% |
| ≥ 35 | | 4.5% |
| **Chronology\*\*** | Prospective study | | Prospective study | | |
| **Reference-standard test** | SCID | | N/A | | |
|  | **EPDS** | **PDSS** | **EPDS** | **PDSS** | |
| **Cut-off (mPPD)** | >9 | >60 | >9 | >60 | |
| **AUC (mPPD)** | 0.77 | 0.87 | N/A | N/A | |
| **Cut-off (MPPD)** | 12/13 | 79/80 | 12/13 | 79/80 | |
| **AUC (MPPD)** | 0.96 | 0.98 | 0.898\* | 0.983\* | |
| **AUC (MPPD & mPPD)** | 0.83\* | 0.91\* | 0.822\* | 0.979\* | |
| **RI administration delay\*\*\*** | None | | N/A | | |
| **Interviewer** | Nurse psychotherapist  (n = unknown) | | N/A | | |
| **Blinding** | Yes (interviewer blind to scores, blinding to clinical information not specified) | | N/A | | |
| *\* EPDS vs. PDSS statistically significant (p < 0.001)*  *\*\* Prospective: Demographic information collected before administration of tests*  *\*\*\* Time delay between administration of index test and reference test*  *N/A: Not available* | | | | | |

The sample size of Beck et al. is smaller than that of Zhao et al., 150 and 842, respectively.

Neither study has performed power-calculations to estimate required sample sizes for statistical significance, nor have they presented arguments for their inclusion criteria. This results in recruitment based on convenience rather than for a specific demographic.

Beck et al. include only women that are 2-12 weeks post-partum and delivered a healthy infant. Zhao et al. include only women that are antepartum and have had obstetric complications.

The population ages and educations of the two articles are comparable.

Beck et al. examine mostly white women whereas Zhao et al. examine Asians. This is also represented in the languages spoken.

The women participating in Beck et al. are 5.6 ± 1.52 (mean ± SD) weeks post-partum whereas the women in Zhao et al. are all in varying gestational weeks antepartum.

## 3.3 Protocols

Both studies collect demographic information before administration of questionnaires and are therefore prospective.

Beck et al. employ the SCID as their reference-standard test. Zhao et al. employ no reference-standard test.

Both tests employ the same cut-offs.

## 3.4 Findings

The PDSS has a larger ROC in all the results, but statistical significance is only reached in a subset of results.

The number of interviewers is unknown in Beck et al.

# 4. Discussion

For a comparison of articles to make sense, the methodology of the articles must be adequately similar.

## 4.1 Test protocol

### 4.1.1 Construct

#### 4.1.1.1 Construct similarity

Beck et al. examine postpartum whereas Zhao et al. examine antepartum. In the DSM-V, depressive disorders can be appended the qualifier ‘with peripartum onset’ if manifestation is during pregnancy or in the 4 weeks following birth1. Following this example, ante- and postpartum depression are not examined as two separate constructs in this thesis.

The PDSS is made specifically for post-partum depression, as seen in some of its questions: “I had trouble sleeping even when my baby was asleep.”7 This question makes no sense in the context of antepartum depression, Presumably, Zhao et al. must have modified this question. No such information is given in Zhao et al.

Beck et al. published their article before the publishing of the DSM-V. A natural concern is that their diagnostic criteria for depression would be different than the ones of Zhao et al. However, both articles use a semi-structured interview with the diagnostic criteria of the DSM-IV as their reference-standard. Secondly, the DSM-V has seen no modifications in the criteria for depression relevant to this thesis, except that the specifier ‘with peripartum onset’ has been added2.

To imply unity among researchers around the criteria of peripartum depression would be excessive, but the working construct of this thesis and the included articles is the construct of the DSM-V.

### 4.1.2 Index test

#### 4.1.2.1 Index test comparability

The comparability of the English and Chinese version of the PDSS is ensured by proper forward-backward translatability and validation9. The same holds true for the EPDS10.

#### 4.1.2.3 Chronology

Both articles collect demographic information before administration of tests. This ensures that test-results do not affect participant reporting of demographic information.

### 4.1.3 Reference standard

For an analysis of a screening tool to be meaningful, a suitable reference-standard test must be used.

In the case of depression, the accepted reference-standard in the literature is a DSM-structured diagnostic interview. A discussion of whether this choice is valid is outside the scope of this thesis.

#### 4.1.3.1 Inter-observer variation

The interview appears to have sufficient interrater reliability with Cohen’s kappas between .7 and 1 for each dimension6,11. An assessment of inter-observer variation for the present observers would have strengthened the results of the studies.

In Beck et al. the interviewer is a nurse psychotherapist. There is no explicit information on whether multiple therapists are used.

#### 4.1.3.2 Diagnostic review bias

Every interviewer can affect the interview differently. It is therefore vital that the interviewer is blind to the scores of the screening test and, in the case of multiple interviewers, their concordance is assessed. If the interviewer preferentially diagnoses patients as depressed if they scored highly on one test, this test’s sensitivity and specificity will be artificially inflated.

In Beck et al., the interviewer is blind to the screening results and can therefore not affect the comparison of the questionnaires.

Whether the interviewer is blind to clinical information is not specified. There is therefore a risk of bias.

#### 4.1.3.3 Diagnostic test comparability

The comparability of the Chinese and English versions of the DSM-IV interview is not sufficiently accounted for. The major study validating the translation contain very few cases of depression12. This weakens a comparison of the studies.

To estimate sensitivity and specificity, information on both true negatives, true positives, false negatives and false positives must be obtained. To know false positives and negatives, all screening results must be confirmed by a reference-standard test. In Zhao et al. this is not the case, as the reference-standard test has not been administered to all participants:

“Relative to other studies that reported the sensitivities and specificities of the screening measures for postpartum depression in comparison with diagnostic instruments […] the present study combined two depression screening tools […] to determine the efficacy without comparison with any diagnostic (e.g., SCID or DIS) instruments.” (p. 117 bottom left)

While many such methods exist, they all come with different methodological considerations and should be employed when a reference-standard test is not available13. Due to Zhao et al. not reporting which method they have used, any analysis of their statistical methodology is reduced to guess-work, and one must therefore hold the conclusions of the study in very low regard. This highlights that methodological errors can render even highly significant P-values unreliable.

## 4.2 Study populations

In general, the method of recruitment and composition of the participants will only affect the generalisability of the results, not the internal validity.

#### 4.2.1.1 Demographic features

Variation in the cultural manifestations of depression can affect the difference between the results of the two studies if one test’s questions more adequately encompass the dimensions of one culture. A generalized recommendation across cultures can therefore be dubious. Given that both studies favour the PDSS, this effect does not appear to influence the comparison critically.

In Beck et al., the mothers’ mean number of days since delivery was 39 (SD = 10.67). If the distribution is approximately symmetrical around the mean, a large amount of the participants will exceed the 4-week postpartum criterion for the DSM-V specifier of peripartum depression. The reasoning for this cut-off in the DSM-V is not expanded upon, and it’s therefore hard to gauge the severity of this discrepancy. It does, however, weaken the comparability of the study to the working-construct.

81% of the women in Beck et al. held at least a college degree. This number is abnormally high for the US, where the average number is 44% for 25-29 year olds, and 42% for 25 and over14. For this to affect the comparison between the questionnaires, one questionnaire must systematically result in a different result due to the educational level of the study participants. Given that the questionnaires read at a 3rd grade level, this seems unlikely.

#### 4.2.1.2 Disease prevalence

Zhao et al. exclusively examine women with obstetric complications. Obstetric complications are a stressor, but it seems unlikely that they will change the peripartum depression construct in a way that will favour either scale, seeing as neither scale contains questions regarding obstetric complications. It might, however, increase the prevalence of depression and therefore be a source of variation for sensitivity and specificity15. The mode of this association is not known to the author of this thesis, however an association has been found empirically. This variation would not bias a comparison of the two tests.

#### 4.2.1.3 Population size

Neither study has done calculations on the amount of participants required to attain sufficient statistical power. Such methods are readily available16. The articles therefore run the risk of recruiting either too few or too many participants, resulting in insufficient statistical significance or an unnecessarily high cost of information, respectfully. This has manifested itself in only one AUC-comparison reaching statistical significance in Beck et al. The traditional level of significance, P < 0.05, is only a guideline and every AUC in Beck et al. favours the PDSS. The P-values in table 4 trend inversely with the number of subjects in each group, indicating that the number of subjects, rather than an underlying lack of difference in AUC, might be able to explain the P-values. This might indicate that the underlying result is that the PDSS is superior and that a larger sample size would have allowed for the results to attain statistical significance.

## 4.3 Flow and timing

#### 4.3.1.1 Disease progression bias (time difference between index test and reference test)

In Beck et al., the reference-test is administered immediately following the index-test. The risk of disease-progression bias is therefore minimal.

#### 4.3.1.2 Induction of signal by questionnaire

Each questionnaire might affect the mental state of the patient and therefore the patient’s response on the following questionnaire. It is therefore important that the questionnaires be administered in random order. This is the case for both studies, however neither study presents whether this randomization has been successful. This weakens the conclusions of the studies.

## 4.4 Considerations in selection of optimal cut-off value

#### 4.4.1.1 Consequences of a false-positive

The consequence of a false-positive screen in a well-managed hospital is a diagnostic interview that has been argued to impose minimal harm on the patient17. However, a positive screen might affect point-of-care behaviour independently of the diagnostic interview due to inappropriate labelling18. The stress imposed on the women from this is gauged to be minimal. However, given the history of underestimating the consequences of a positive screen in breast-cancer screening, further research is essential.

#### 4.4.1.2 Consequences of a false-negative

A false negative might falsely reassure caregivers that the patient is not depressed. This can lead to a lack of appropriate treatment and therefore worse outcomes for the patient.

Deciding on the optimal cut-off score is therefore not simply a question of finding the point on a ROC-curve closest to the top-left corner, but a consideration of the above trade-offs.

For this thesis’ comparison to be improved upon, inclusion of further studies is needed containing better methodology consistent with the accepted practices and with sufficiently large study populations to warrant conclusions for both major and minor depression. A broader search scope would be relevant for a comprehensive meta-analysis, allowing forest-plots to gauge the risk of publication bias. Such a plot would be meaningless with a study-sample of n=2.

# 5. Conclusion

Both studies indicate that the PDSS is the most precise questionnaire, however, due to the small study-sample and methodological discrepancies between the studies no firm recommendation of either questionnaire can be made. Further research is needed.

# References

1. Association, A. P. & Force, D.-5. T. *Diagnostic and statistical manual of mental disorders : DSM-5.* (dsm.psychiatryonline.org, 2013).

2. Highlights of Changes from DSM-IV-TR to DSM-5. 1–19 (2013).

3. Neiman, S., Carter, S., Van Sell, S. & Kindred, C. Best practice guidelines for the nurse practitioner regarding screening, prevention, and management of postpartum depression. *Crit Care Nurs Q* **33,** 212–218 (2010).

4. Chisholm, D. *et al.* Scaling-up treatment of depression and anxiety: a global return on investment analysis. *Lancet Psychiatry* **3,** 415–424 (2016).

5. Beck, C. T. & Gable, R. K. Postpartum Depression Screening Scale: development and psychometric testing. *Nurs Res* **49,** 272–282 (2000).

6. Maffei, C. *et al.* Interrater reliability and internal consistency of the structured clinical interview for DSM-IV axis II personality disorders (SCID-II), version 2.0. *J. Pers. Disord.* **11,** 279–284 (1997).

7. Beck, C. T. & Gable, R. K. Comparative analysis of the performance of the Postpartum Depression Screening Scale with two other depression instruments. *Nurs Res* **50,** 242–250 (2001).

8. Zhao, Y. *et al.* Combined use of the postpartum depression screening scale (PDSS) and Edinburgh postnatal depression scale (EPDS) to identify antenatal depression among Chinese pregnant women with obstetric complications. *Psychiatry Res* **226,** 113–119 (2015).

9. Li, L., Liu, F., Zhang, H., Wang, L. & Chen, X. Chinese version of the Postpartum Depression Screening Scale: translation and validation. *Nurs Res* **60,** 231–239 (2011).

10. Wang, Y. *et al.* Psychometric evaluation of the Mainland Chinese version of the Edinburgh Postnatal Depression Scale. *Int J Nurs Stud* **46,** 813–823 (2009).

11. What is the relaibility of the SCID-II? *scid.org* Available at: http://www.scid4.org/psychometric/scidII\_reliability.html. (Accessed: 26 October 2016)

12. So, E. *et al.* The Chinese-bilingual SCID-I/P Project: Stage 1 — Reliability for Mood Disorders and Schizophrenia. *Hong Kong Journal of Psychiatry* 7–18 (2003).

13. Reitsma, J. B., Rutjes, A. W. S., Khan, K. S., Coomarasamy, A. & Bossuyt, P. M. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *J Clin Epidemiol* **62,** 797–806 (2009).

14. **Educational Attainment in the United States: 2014**. Available at: https://www.census.gov/hhes/socdemo/education/data/cps/2014/tables.html. (Accessed: 30 October 2016)

15. Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S.QUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J Clin Epidemiol* **66,** 1093–1104 (2013).

16. Hess, A. S. *et al.* Methods and recommendations for evaluating and reporting a new diagnostic test. *Eur. J. Clin. Microbiol. Infect. Dis.* **31,** 2111–2116 (2012).

17. Gjerdingen, D. K. & Yawn, B. P. Postpartum depression screening: importance, methods, barriers, and recommendations for practice. *J Am Board Fam Med* **20,** 280–288 (2007).

18. Sheehan, A. M. & McGee, H. Screening for depression in medical research: ethical challenges and recommendations. *BMC Med Ethics* **14,** 4 (2013).