Author's response to reviews

Title: Reliability and validity of the Thai version of the PHQ-9

Authors:

   Manote Lotrakul (ramlt@mahidol.ac.th)
   Sutida Sumrithe (rasda@mahidol.ac.th)
   Ratana Saipanish (rarsp@mahidol.ac.th)

Version: 3 Date: 18 February 2008

Author's response to reviews: see over
Re: MS: 1603880518160237 – “Reliability and validity of the Thai version of the PHQ-9”

Dear the Editor:

Thank you very much for giving us the opportunity to revise our manuscript. We would like to thank the reviewers’ for giving us the highly valuable comments to improve our manuscript. We have revised the manuscript in accordance with all the reviewers’ comments. A point-by-point response to the comments is given below. We hope that the revised manuscript and our responses are satisfactory to you.

Thank you for your consideration. We look forward to hearing your reply.

Sincerely yours,

Manote Lotrakul, M.D.
Reviewer 1: Yvonne Forsell

General

1. The paper presents a carefully designed study concerning the validity and reliability of the Thai version of the PHQ-9. It is clearly written and easy to understand. However it needs some editing of the English since there are several grammatical errors as well as spelling errors.

Response: The final manuscript will be sent to a native English speaking colleague for language revision.

2. The number of depressed is surprisingly low (1.9%) but as the author states this could have been due to a selection bias. This would mean that people who were depressed had chosen to go to the psychiatrists instead of to the GPs. However it could also mean that the PHQ-9 does not capture depression due to underreports of symptoms.

Response: We agree with the reviewer. We have included this suggestion in the discussion section on limitations of the study – “Second, there was a possibility that a proportion of participants might have underreported their depressive symptoms on both the PHQ-9 and the MINI clinical interview.” (page 15, 3rd paragraph)

Major Compulsory Revisions

1. The statement that PHQ-9 was the optimal measure to establish a categorical diagnosis of depression is not supported. The authors might mean when comparing to PHQ-9 used as a continuous measure.

Response: We thank the reviewer for her thoughtful comment. This section has been revised to provide more clarity. Result of the PHQ-9 categorical algorithm for detecting major depression showed that it is less useful than the cut-off score for a screening purpose.

2. A ROC curve is presented and the area under the curve is 0.89. The authors state that this means that the PHQ-9 is a good test to identify patients with depressive disorder. This is not entirely true, a better way to analyse the performance of PHQ-9 would be to use QROC curves.

Response: We have revised the statement about the AUC value as “The area under the curve (AUC) in this study was 0.89 (SD=0.05, 95% CI 0.85 to 0.92) which demonstrated a moderate accuracy [21].” We thank the reviewer for her kind suggestion of the QROC curves. We agree that this is a better way to analyze the performance of PHQ-9. However, this involves a somewhat sophisticated statistical analysis and a software which is difficult for us to access. As most articles so far report the performance of a screening tool using the ROC curve, we hope that the reviewer accept our request to retain the original ROC curve for analyzing the performance of the PHQ-9.

---

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

1. In the introduction it is stated that people are reluctant to seek help from a psychiatrist since it implies that they have psychiatric disorders. It is natural that you suspect that you have a psychiatric disorder if you go to a psychiatrist. What the authors intends to say is probably that people are reluctant to seek due to prejudices against psychiatric disorders. This has been reported in studies from several countries. Thus GPs are important all over the world not only in Thailand.

Response: This section has been revised as suggested - “In Thailand, most patients with mental disorders are often treated by general practitioners (GPs) because there are still a small number of psychiatrists in the country. Moreover, people are reluctant to seek help form a psychiatrists due to prejudices against psychiatric disorders. This has been reported in studies from several countries [3-5]. Thus, GPs in Thailand, as in many other countries, play an important role in taking care of patients with mental heath problems.” (page 4, 1st paragraph)

2. In the methods section it is stated on page 7 that the Thai version of MINI had a high Kappa, sensitivity and positive predictive value. The numbers in the parenthesis is confusing, which is the Kappa value? Which diagnostic procedure was the golden standard?

Response: This section has been revised to make it clearer – “The MINI, version 5, is a standardized clinical diagnostic interview schedule for DSM-IV Axis-I disorders [19]. It can be reliably administered by lay interviewers using appropriate training. The MINI, Thai version which was translated from the MINI, version 5 was use in this study. A study of the MINI, Thai version comparing
with a clinician interview showed that it had a high sensitivity of 0.92 and specificity of 0.93 on the diagnosis of current major depressive episode [16]. The depression modules of the schedule were used in the study as the gold standard diagnostic tool.” (page 8, 2nd paragraph)

3. Table 3: Less than major depression could be rephrased to moderate depression, which is a more commonly used term.
Response: We have changed the wording as suggested.

4. The term fiscal usually refers to yearly finance and maybe it would be better to use the word financial.
Response: We have changed the wording as suggested. –“Financial and institutional constraints in health care services should be taken into account before adopting such service delivery programs in order to maintain successful care” (page 16, 2nd paragraph)
Reviewer 2: Antonio Lobo

Major Compulsory Revisions

1. Clarity of the manuscript is an issue. For example, “five consecutive patients were invited…” (page 6, 2nd paragraph); this should be clarified.
Response: We have changed this sentence to “Every fifth patients attending the outpatient clinic of the department was invited to participate in the study while they were waiting for consultation.” (page 6, 2nd paragraph)

2. It is not clear how the diagnosis of major depression was performed (page 7, 3rd paragraph and page 10, 2nd paragraph).
Response: We have revised the methods section: measures to make it clearer – “We used the Mini International Neuropsychiatric Interview (MINI), Thai version [16] as the criterion standard for the presence of major depressive disorder.” (page 6, last paragraph).

3. “Validity analysis” (page 10, 2nd paragraph). The data showing that mean scores in PHQ-9 are higher in “major depressions” than in non-depressed patients, it is not enough to support the validity.
Response: This section has been revised to provide more clarity in the methods: data analysis section – “For validity analysis, we determined both criterion and convergent validity. Criterion validity tests a scale’s performance in comparison to a gold standard. The MINI, Thai version diagnosis of major depressive disorder was used as a criterion standard. Convergent validity is present when a scale behaves according to hypothesized relationships between two tests that presumably measure the same construct. In this study we tested the association between PHQ-9 and the HAM-D.” (page 8, last paragraph and page 9, 1st paragraph)

4. “PHQ-9 as a criterion-based measure” (page 10, 3rd paragraph). The authors should clarify this paragraph. It is not clear what is the “gold standard” to document the sensitivity, etc., of the PHQ-9.
Response: In fact, we intended to mean - “PHQ-9 as a criterion-base measure by using categorical algorithm approach”. To avoid this ambiguity, we have revised this section and this subheading was deleted. The gold standard in this study is the MINI, Thai version.

5. The sensitivity documented in this paragraph (0.53) is low for screening instruments, I believe the comments in the discussion section are insufficient.
Response: The value (0.53) mentioned above is the sensitivity of the PHQ-9 using a categorical algorithm. This approach yielded a low sensitivity and high specificity makes it not suitable for a screening purpose. We have added in the discussion section as – “Result of the PHQ-9 categorical algorithm for detecting major depressive episode showed that it provided very high specificity. However, its sensitivity was poor (0.53), rendering this algorithm less useful than the cut-off score for a screening purpose. On the other hand, its high positive likelihood ratio (LR+) of 27.4 may make it a suitable method for a diagnostic purpose” (page 11, 2nd paragraph)

6. PHQ-9 as a severity measure (page 11, 1st paragraph). What is the gold standard to calculate the ROC curve here?
Response: We used the MINI, Thai version as the gold standard to calculate the ROC curve. We have added a sentence – “To determine the best cut-off score, the receiver operating characteristic (ROC) curve was constructed against the presence of major depressive disorder by the MINI.” (page 9, 2nd paragraph). The title for Figure 1 was changed to – “Figure 1: The Receiver Operating Characteristic (ROC) curve of the PHQ-9 versus the MINI for major depression diagnosis”

Response to the notion 3-6: We apologize for this confusion. As proposed by Kroenke et al [Kroenke K, Spitzer RL, Williams JB: The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001, 16:606-613.], there are two approaches for using the PHQ-9.
- It can be used to establish depressive disorder diagnosis using a categorical algorithm. Major depressive disorder is diagnosed if 5 or more of the 9 symptoms have been present at least more than half the days of the past 2 weeks and 1 of these symptoms is either depressed mood or anhedonia.
- It can also be used as a continuous measure for the screening purpose with recommended cut-off scores for the diagnosis of major depression.

We tested the validity of both methods of the PHQ-9. This detail has been included in the methods: measures section (page 7, last paragraph)
7. A positive predictive value of 0.21 is quite low (page 11, 1st paragraph).

Response: We agree with the reviewer. We have addressed this in the discussion section in the revised manuscript – “The cut-off score at this point yielded the low PPV (0.21) and high NPV (0.99). Results from other studies of the PHQ-9 showed PPV from 0.31 to 0.51 (depending on the cut-off) [11]. The PPV is the probability of disease if the patient’s test is positive, while NPV indicates the probability that the patient is disease-free if the test is negative. A low PPV may indicate lower specificity, lower disease prevalence in the population undergoing screening, or a combination of these factors [29]. In this study, results may come from both low specificity (0.77) and low disease prevalence (6.8%).” (page 15, 1st paragraph).

8. On the bases of these comments, I do not feel satisfied that the conclusions in the Abstract are well substantiated. Furthermore, I think there are no data to support the statement that this version of the PHQ-9 “could be a valuable tool for … and monitoring of depression in this population”.

Response: We have revised the conclusions in the abstract as “The Thai version of the PHQ-9 has acceptable psychometric properties in screening for major depression in general practice with a recommended cut-off score of nine or greater”.

We have also revised the summary as “In summary, the Thai version of the PHQ-9 has acceptable psychometric properties in screening for major depression in general practice with a recommended cut-off score of nine or greater. Due to the low PPV in this study, a further clinical assessment is recommended if a result of the scale is positive. The categorical algorithm of the PHQ-9 yielded low sensitivity makes it less suitable for the screening purpose.” (page 16, last paragraph)

Minor Essential Revisions
The manuscript would improve if the Methods section is clearly distinguished from the Results section.
Response: We thank the reviewer for his thoughtful suggestion. We have modified these sections as per this suggestion.

Discretionary Revisions (which the author can choose to ignore)

Quality of written English: Needs some language corrections before being published
Response: The final manuscript will be sent to a native English speaking colleague for language revision.
Reviewer 3: Abiodun Adewuya

Major Compulsory Revisions

1. not enough argument for the use and validation of PHQ-9. The authors should talk about the brevity of the instrument and its strict adherence to DSM-IV criteria for depression. The age mentioned may not be so convincing since older instruments like BDI now have BDI 2 etc.

**Response:** We thank the reviewer for his valuable suggestion. The main reason for us in choosing to translate and validate the PHQ-9 was because of its brevity. Most of self-rating questionnaires in Thailand consist of 15-20 items which are too burdensome for routine use in primary care. We have added sentences as suggested. - “At only 9 items, the PHQ-9 is substantially shorter than most depression screening measures. It comprises the 9 diagnostic symptom criteria upon which the diagnosis of DSM-IV major depressive disorder is based. The PHQ-9 has the potential of being the dual-purpose instrument that, with the same 9 items, can establish depressive disorder diagnoses using a categorical algorithm as well as grade the depressive symptom severity” (page 5, 2nd paragraph)

2. What is the point in screening when there is no programme and infrastructure in place for treatment?. The authors need to marshall their points very well.

**Response:** We agree with the reviewer. We have addressed this in the discussion section in the revised manuscript. – “Caution should be exercised that although the PHQ-9 is a valid and useful self-rating instrument in screening for depression, the routine usage of the questionnaire, in isolation, may not improve the quality of care as anticipated. Results from a review has revealed that only routine employing of screening questionnaires for depression has minimal impact on clinician’s detection and management or outcome of depression [33]. Moreover, in developing countries, new psychiatric cases found from such instruments may just be the add-on burden to GPs whose workload has been heavy. Financial and institutional constraints in health care services should be taken into account before adopting such service delivery programs in order to maintain successful care [34]. In a region where there is a shortage of GPs, besides the benefit of case detecting tool, a structured approach that facilitates an increased role for non-medical staff, patients, and family members may be more appropriate [35].” (page 16, 2nd paragraph)

3. What is the time frame of the study?, Years?

**Response:** We have added a sentence under the methods section as - “The patients were recruited between October 2006 and February 2007 from the outpatient clinic of the department of family medicine, Ramathibodi Hospital, Bangkok.” (page 6, 1st paragraph)

4. “...first of 5 consecutive patients invited...”.. how then was 1000 cases selected?, from 5000 patients?

**Response:** We have reworking as “Every fifth patients attending the outpatient clinic of the department was invited to participate in the study while they were waiting for consultation” (page 6, 2nd paragraph)

Our samples were recruited from approximately 5,000 patients.

5. How was the random assessments made

**Response:** The patients were randomly assessed by using convenience sampling. Our outpatient department is one of the busiest clinics in the hospital. Due to time constrain, patients were selected because their convenient accessibility to the researcher. We have added this in the methods section as- “After completing the questionnaire, the patients were then randomly assessed by using convenience sampling. A research assistant who was unaware of the patients’ PHQ-9 results interviewed the patients until a total of 300 responses was attained.” (page 6, 3rd paragraph)

6. translational method not well explained

**Response:** We have revised this section as - “After obtaining permission from the copy right holder, the PHQ-9 was translated following the guidelines for cross-cultural adaptation of self-report measures [18]. The process included two independent forward translations of the original PHQ-9 into Thai, consensus between translators on a forward translation, back-translation by a bilingual English teacher, and a review of the back-translation. Ten patients attending the out-patient department were invited to complete and give comment on the prefinal version. Final modifications and adjustments were made accordingly.” (page 7, 2nd paragraph)

7. How different was the randomly selected 300 from the rest?

**Response:** The sociodemographic characteristics of respondents who were further interviewed were not different from the first group. The mean PHQ-9 score for 924 subjects was 4.93 (SD=3.75) whereas the mean score for the 279 subjects was higher (6.5, SD=4.29). Reliability analysis of the interviewed
group yielded the Cronbach's alpha of 0.81 which is not different form the first group. Item analysis of both groups showed the same item profile. (page 10, 2nd and 3rd paragraph)

8. in what language was MINI administered?, if the PHQ-9 was in Thai, why not the MINI?
Response: The MINI, Thai version was conducted in Thai. Thai is the national and official language of our country. Although there are a few regional dialects, most of Thai people can use the official Thai well.

9. the limitations of the MINI was not discussed, it is really a highly structured instrument that it may even be seen as a screening instrument instead of a diagnostic instrument
Response: We have added the limitation of MINI as “Third, although the MINI, Thai version performed well in a validity study, due to its highly structure instrument, it is still possible that it overestimated or underestimated the rate of depression in this study. However, we found nearly the same as the study of Adewuya et al [32] that the MINI is structured in simple, lay language, that was easy to understand and few problems were encountered in the administering the instrument.” (page 15, last paragraph and page 16, 1st paragraph)

10. What of mild depression -- the criteria by DSM-IV should be discussed
Response: In our revised manuscript we did not mention of mild depression as we would like to focus our presentation only on major depressive disorder.

11. I think Adewuya et al (2007) published in Journal of affective Disorder should be a useful addition tpto the reference list
Response: We thank the reviewer for his suggestion. The mentioned article (Adewuya et al. Impact of postnatal depression on infants' growth in Nigeria. J Affect Disord. 2007 Nov 6;) may not be relevant to our manuscript. We have added articles from Adewuya to the reference list as below;


Discretionary Revisions (which the author can choose to ignore)

Quality of written English: Needs some language corrections before being published
Response: The final manuscript will be sent to a native English speaking colleague for language revision.
Reviewer 4: Janet Williams

Minor Essential Revisions

1. On page 6, paragraph two, the authors state “The first of five consecutive patients were invited to complete the PHQ-9…” It would be helpful to clarify if this means that every fifth patient entering the hospital was invited to participate. A subheading for assessments used may also be helpful.

Response: We have reworded as suggested. - “Every fifth patients attending the outpatient clinic of the department was invited to participate in the study while they were waiting for consultation.”

A subheading of the methods section was added.

2. On page 9, the first paragraph under the heading “Reliability and item analysis” would be more appropriate under the methods section as it describes the process and problems encountered during translation.

Response: We agree with the reviewer. We have moved this paragraph to the methods section.

3. On page 10, the paragraph under validity analysis is a bit unclear. It appears that this paragraph is referring to the use of the MINI to determine the validity of the PHQ-9. However, this needs to be stated more clearly and in more detail.

Response: We have revised this section and added more details on validity analysis to make it clearer.

- “For validity analysis, we determined both criterion and convergent validity. Criterion validity tests a scale’s performance in comparison to a gold standard. The MINI, Thai version diagnosis of major depressive disorder was used as a criterion standard. Convergent validity is present when a scale behaves according to hypothesized relationships between two tests that presumably measure the same construct. In this study we tested the association between PHQ-9 and the HAM-D.” (page 8, last paragraph and page 9, 1st paragraph)

4. On page 11, the paragraph under “PHQ-9 as a severity measure” reports sensitivity, specificity, and predictive values using a cutoff score of > 9. However, when examining the referenced table (Table 2), the reported numbers correspond to a cutoff score of “8/9.” This is a bit confusing and needs to be clarified. The authors should also describe table 3 in more detail. They indicate that the ANOVA revealed a significant difference but do not expand on what this means. An additional sentence or two related to table 3 would be helpful.

Response: We have change the text to “using a cutoff score of nine or greater” (page 12, last paragraph). We have also changed the mathematical symbol in Table 2 to avoid this problem.

- “We divided HAM-D score into 4 groups according to the severity of depression [22]. We hypothesized that subjects with major depression would have higher PHQ-9 scores compared to subjects with mild depression and no depression. As shown in the Table 3, the group of patients with major depression by the HAM-D had a mean PHQ score of 14, followed by a mean PHQ score of 10.05 and 8.14 in patients with moderate depression and mild depression, respectively. These differences among four PHQ-9 severity subscales were statistically significant (ANOVA: F=31.91, df=3, 275, p<0.0001).” (page 12, 3rd paragraph)

5.1 On page 14, the last sentence of the first paragraph that begins “this finding demonstrates that in a society…” is awkward and unclear. It seems that the authors are trying to state that their results refute the notion that subjects from countries that discourage expression of emotional distress report more somatic symptoms than subjects from countries where expression of emotional distress is accepted or even encouraged.

Response: We have revised this section as suggested. – “This finding results support Kirmayer’s argument [26] that subjects from countries that discourage expression of emotional distress do not report more somatic symptoms than subjects from countries where expression of emotional distress is accepted or even encouraged.” (page 14, 1st paragraph)

5.2 Also, in the last paragraph on page 14, the authors indicate that high specificity is more useful due to the high patient load of GPs. The authors do not discuss the potential consequences of missing a case due to this increased specificity. They might discuss the feasibility of perhaps allowing a lower threshold on the PHQ-9 to trigger further assessment by a mental health professional or some other follow up which can refine the patient pool before the GP becomes involved.

Response: We thank the reviewer for her thoughtful suggestion. After a lengthy discussion, we have decided to omit this section because the argument we proposed was not solid enough.
6. On page 15 in the paragraph that begins “Our study had several limitations”, the authors state that the “main objective of the study was to assess the ability of the tool in identifying borderline cases of depression…” This is the first time that this specific objective is mentioned. It seems that this should have been expressed earlier in the paper.

Response: The specific objectives of the study were added in the introduction section as per this suggestion.

7. Lastly, on page 16 the second sentence in the paragraph that begins “caution should be exercised…” states, “The results from the review…” Upon first inspection, it is unclear what review is being referred to here. It would be helpful to simply change the sentence to “The results from a review…”

Response: We have corrected as per this suggestion.

We thank the reviewer very much for her thoughtful comments and clear suggestions that have helped us to improve our manuscript considerably.

Discretionary Revisions (which the author can choose to ignore)

Quality of written English: Needs some language corrections before being published
Response: The final manuscript will be sent to a native English speaking colleague for language revision.