Reviewer's report

Title: Standardization of the Colombian Version of the PHQ-4 in the general population

Version: 2 Date: 11 June 2014

Reviewer: Kurt Kroenke

Reviewer's report:

This is a useful paper providing important psychometric data on the PHQ-4. In addition to the fact it provides data for Columbia, it also provides data on the PHQ-4 in a large population sample. While there is much data on the PHQ-9 and GAD-7, there is less data on the 2 brief depression and anxiety screeners that are part of the longer measures – the PHQ-2 and GAD-2, which together form the PHQ-4. Thus, this paper makes an important contribution to the literature on the PHQ-4, a public domain increasingly used depression and anxiety screener.

Minor Essential Revisions

1. Table 1, the n² in last column is not explained anywhere in the paper. It should be explained in a footnote to Table 1, and also in the Data analysis section. It should also be explained how to interpret this statistic. It appears that Cohen’s d is used for binary variables, and this n² is used where there are more than 2 subgroups, but this is not explained. Is it a type of effect size? And does one interpret the magnitude of a n² similar to Cohen’s d?

2. Related to #2 above, the authors state on p. 7: “As noted in Table 1, the calculated effect sizes were low for gender and household income, and moderate for unemployment and education.” The authors should indicate what values of Cohen’s d and n² represent “low” and “moderate” effect sizes. For Cohen’s d, values of .2, .5, and .8 are sometimes considered small, medium, and moderate effect sizes (Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989; 27:S178-S189). The authors might have a different citation but they should define what they mean by “low” and “moderate” and provide a citation. Also, when they say it was “low” for gender and “moderate” for unemployment, the values are similar (.14 and .12) though one is Cohen’s d and the other is n². Thus, please provide an explanation.

3. In the data analysis section, it is stated that a good fit by the CMIN/DF ration should be close to 3 or smaller. In the Results, it is stated that all fit indices were good except the RMSEA. However, the CMIN/DF ratio was 32.31, which is nowhere near 3. This would suggest both the CMIN/DF and RMSEA indices did not meet criteria for good fit. Please reconcile.

4. The tests for invariance in the Data Analysis section, Results, and Table 3 do not entirely make sense and are confusing.
a. Data analysis section provides a rather detailed description as follows: Measurement invariance was tested in three steps using first the configural, combined model (no constraints), followed by a metric invariant model (with equal item loadings, that is paths and covariances constrained equal), and a scalar invariant model (with equal item loadings and item intercepts across groups) [29]. Since these models are hierarchically nested and increasingly restricted, the models were then compared to each other on the basis of the 2 differences \( \Delta \text{CFI} \) and \( \Delta \text{RMSEA} \). Values \( \Delta < 0.01 \) indicate invariance of the model [30].

b. However, Table 3 does not seem to match the above 3 steps. Moreover, we do not see anything about \( \Delta \text{CFI} \) and \( \Delta \text{RMSEA} \). Thus, for a reader unfamiliar with these tests of invariance, there seems to be little correspondence between the analyses described in the Analysis section and what is shown in Table 3.

c. What is described then in Results is also confusing. It is stated: “Results indicate that the two-factor model was structurally invariant between age and gender groups. According to Cheung (2009), only p values \( \Delta < 0.01 \) indicate the invariance of the models [30]; that is, the p-level of 0.05 was considered invariant. The results are NS for gender and 0.05 for age, which are NOT less than 0.01. So how does this suggest invariance since they are greater than 0.01?"

d. Also, there should be a sentence somewhere stating in lay terms what “invariance” means. Does that mean the 2-dimensional model performs similarly across gender and age? Or does it mean something else?

5. Another study limitation is that no data is available on the nearly one-third of contacted individuals who did not participate (to know if they are similar to respondents) so it is possible there is some selection bias in the sample. This should be noted in Discussion.

6. Minor point. The abbreviation used for the Questions on Life Satisfaction scale is FLZ, which does make sense to the English reader. Where does FLZ come from? Should the abbreviation instead be QLS?

7. On p. 8, it is stated: “Discriminant validity can be assumed in terms of significant intercorrelations of the PHQ-4 with self-efficacy \( (r=-0.26, p<0.001) \) and life satisfaction \( (r=-0.29, p<0.001) \). What do the authors mean by “discriminant” validity? Do they mean the correlations are less than those for the pure mental measures (HADS and GHQ)? If so, do they mean “divergent” validity, in that correlations, while significant, are less than that for the purer mental measures? Or do they mean something else by “discriminant”?

8. Related to #9 above, on p. 9, it is stated: “Correlations with the PHQ-4 were highest for the total score of the Hospital Anxiety and Depression Scale, although only marginally higher than for the discriminating other scales.” First, the grammar “discriminating other scales” is poor. More importantly, the term “discriminating” needs to be clarified or replaced. And the differences in correlations with the HADS and GHQ compared to the other scales does not look marginally different but indeed moderately larger.
9. On p. 10, it is stated: “Though a review of randomized controlled trials with implementation of screening for depression and anxiety symptoms in routine care revealed little or no impact on the recognition, management, or outcome of depression in primary care or the general hospital [40] …” This review only focused on depression and not on anxiety, so anxiety should be deleted from the first part of sentence.

Discretionary Revisions

10. The Introduction is fairly long. One optional paragraph is the second paragraph on p. 4, lines 6 through 14, which could be deleted.

11. Table 2. A footnote to the first half of the table might indicate items 1 and 2 are the PHQ-2 depression items, and items 3 and 4 are the GAD-2 anxiety items. Also, a footnote should indicate what level of significance is indicated by the asterisks.

Level of interest: An article of importance in its field

Quality of written English: Needs some language corrections before being published

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests.