Author's response to reviews

Title: Psychometric properties of the IDS-SR30 for the assessment of depressive symptoms in spanish population

Authors:

Margalida Gili (mgili@uib.es)
Juan V Luciano (jvluciano@pssjd.org)
Natalia Bauzá (nataliasiddons@hotmail.com)
Jaume Aguado (jvluciano@pssjd.org)
María J Serrano (mj.serrano@uib.es)
Silvia Armengol (silvia.armengol@almirall.com)
Miquel Roca (mroca@uib.es)

Version: 2 Date: 1 August 2011

Author's response to reviews: see over
REVIEWER 1 comment (Ira Bernstein)

*I am not convinced that the third factor is worth the loss in parsimony. This can be addressed by reporting the Akaike Information Criteria (or similar) statistics that are reported in their MPlus output. Other than that, I think the paper makes a valuable contribution. I consider this a minor essential revision.*

RESPONSE TO REVIEWER 1

We greatly appreciate the reviewer kind and encouraging comments about the quality of our manuscript. We have followed his suggestion, trying to incorporate it into the revised version of our manuscript, as follows:

- Firstly, we have included the AIC value of the five models in Table 4.

Table 4. CFA of five factor models of the IDS-SR\textsubscript{30} in the follow-up dataset (n= 1594).

<table>
<thead>
<tr>
<th>Model</th>
<th>Source</th>
<th>Df</th>
<th>$\chi^2$</th>
<th>TLI</th>
<th>CFI</th>
<th>RMSEA (90% CI)</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-factor</td>
<td>e.a. Trivedi et al. (2004)</td>
<td>350</td>
<td>3793.5</td>
<td>0.93</td>
<td>0.94</td>
<td>0.079 (0.076-0.081)</td>
<td>70462.8</td>
</tr>
<tr>
<td>2-factor</td>
<td>Bernstein et al. (2006)</td>
<td>347</td>
<td>3692.2</td>
<td>0.93</td>
<td>0.94</td>
<td>0.078 (0.076-0.080)</td>
<td>70339.9</td>
</tr>
<tr>
<td>3-factor</td>
<td>Rush et al. (1996)</td>
<td>342</td>
<td>3214.3</td>
<td>0.94</td>
<td>0.95</td>
<td>0.073 (0.070-0.075)</td>
<td>70018.2</td>
</tr>
<tr>
<td>3-factor</td>
<td>Wardenaar et al. (2010)</td>
<td>344</td>
<td>2774.2</td>
<td>0.95</td>
<td>0.96</td>
<td>0.067 (0.064-0.069)</td>
<td>69926.9</td>
</tr>
<tr>
<td>3-factor</td>
<td>PCA, present study</td>
<td>344</td>
<td>3245.4</td>
<td>0.94</td>
<td>0.95</td>
<td>0.073 (0.070-0.075)</td>
<td>69928.3</td>
</tr>
</tbody>
</table>

Note: IDS-SR\textsubscript{30} = Inventory of Depressive Symptomatology Self Report; Df = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; 90% CI = 90% confidence interval of the RMSEA. AIC = Akaike Information Criteria. Indices that indicate the best model fit are printed in bold font.

- Secondly, we have slightly modified two paragraphs of the manuscript in order to include information related to the AIC (see text in bold):

Page 9. Although a model with a non significant chi-square estimate is generally considered a good fitting model, Hu and Bentler recommended combinational rules to evaluate model fit [38]. Therefore, the following indices were analysed (values in parentheses denote goodness-of-fit standards): the Tucker-Lewis, non-normed fit index (TLI\geq 0.95), the comparative fit index (CFI\geq 0.95), and the root means square error of approximation (RMSEAs\leq 0.08). We also report the Akaike’s Information Criterion (AIC), a relative fit index especially designed for comparing alternative factor models. The model with the smallest AIC value has the best fit. Used together, these indices provide a more conservative and reliable evaluation of the solution.
Fit statistics for the factor models are shown in Table 4. Although all models showed RMSEA values that are considered acceptable (< 0.08), the one-factor and two-factor solutions failed to provide good fit taking the other indices into account. Considering the findings collectively, all three-factor models seem adequate for the current sample. However, as the Hu and Bentler’s guidelines for retaining a hypothesized model recommended, the factor structure posited by Wardenaar [32] received more support than the other three-factor structures because obtained the lowest AIC value and its CFI and TLI values reached the rule of thumb minimum value of 0.95. Furthermore, the TLI value was 0.01 greater (0.95 vs. 0.94).
REVIEWER 2 comments (Edwin de Beurs) and responses.
Major Compulsory Revisions

1. The abstract is too long and offers rather detailed information, which belongs to the main body of the text.

We have reduced the abstract and eliminated detailed information that now appears only in the main body of the text

2. In the abstract (and again in the method section of the manuscript) the HDRS is presented as “the gold standard”. There is no need to do so: the HDRS can (and should) be presented as the measure used to assess convergent validity. Twice it is mentioned that patients had been in treatment for at least 6 weeks. Why was this a requirement? And it makes me wonder whether the first assessment can truly be considered a baseline assessment as the baseline was at least 6 weeks earlier. Some clarification should be offered (in the text, not in the abstract).

Following the suggestions of the reviewer, we have removed the references to the HDRS as a gold standard in the abstract and the section method. It is now presented as the measure that was used to assess convergent validity.

As the reviewer correctly states, the first evaluation is not really a baseline. We have therefore replaced the term “baseline” for “first evaluation” throughout the text. Because the RESIST study, as discussed in the article, was designed to evaluate the residual symptoms of depression, it was required that patients had been treated with antidepressant at least 6 weeks to have depressive patients in remission at first evaluation. This comment has been added to the text in “procedure” in the method section.

3. On page 10 the authors present means for baseline and follow-up scores for the IDS. The baseline score could be included in Table 1 (accompanied by the pretest scores on the HDRS or the proportion of patients scoring beyond the cut-off for the HDRS, as this information is essential in describing the sample. The findings regarding item 4 suggest that a reformulation (“Sleeping more than usual”?) is in order. Is this item also problematic in other language versions?

The scores obtained in IDS-SR$_{30}$ and HDRS in the first assessment have been included in the Table 1. As far as we know, item 4 do not offer problems in other versions.
4. Convergent validity: In Table 6 the correlation between change scores could be added. In addition, the pre and posttest scores on the IDS and the HDRS allow for a comparison of the sensitivity to change of both measures. I suggest an analysis with SPSS GLM for a 2 by 2 design with time and instrument as factors. A significant interaction effect would reveal a difference between both measures in sensitivity to change. Finally, the data allow for an investigation of relevant cut-off points for statistical and clinical significant change. For the HDRS cut-off values are published. Which cut-off value on the IDS yields the best correspondence with HDRS defined outcome of treatment?

Following the reviewer’s suggestion, we have performed two new analyses: (1) As can be seen in Table 6, we have added the correlations between change scores. The correlations were all statistically significant, positive, and large (>0.50). Additionally, some paragraphs in the paper have been modified to comment this new analysis (see text in bold).

Table 6. Correlations between the IDS-SR$_{30}$ (total and sub-scale scores) and the HDRS at baseline and at follow-up.

<table>
<thead>
<tr>
<th>HDRS</th>
<th>Mood/Cognition</th>
<th>Anxiety/Somatic</th>
<th>Sleep</th>
<th>IDS-SR$_{30}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.82*</td>
<td>0.70*</td>
<td>0.70*</td>
<td>0.85*</td>
</tr>
<tr>
<td>Follow-up</td>
<td>0.82*</td>
<td>0.67*</td>
<td>0.64*</td>
<td>0.85*</td>
</tr>
<tr>
<td>Change</td>
<td>0.80*</td>
<td>0.67*</td>
<td>0.66*</td>
<td>0.84*</td>
</tr>
</tbody>
</table>

Note: *Correlation significant at $p<0.001$

Page 9. Construct validity. Pearson product-moment correlations computed between the IDS-SR$_{30}$ factors and the HDRS. Additionally, correlational analyses between change scores in the IDS-SR$_{30}$ and change scores in the HDRS were computed in order to know the external longitudinal construct validity of the IDS-SR$_{30}$. We took the Cohen’s criteria (Cohen, 1988) into account to evaluate the substantive significance of correlations (large correlations are those > 0.50, medium correlations are from 0.30 to 0.49, and small correlations are from 0.10 to 0.29).


The IDS-SR$_{30}$ would show adequate convergent validity if the total and subscale scores correlated significantly with the HDRS. As can be seen in Table 6, the correlations between Mood/Cognition, Anxiety/Somatic, Sleep, the IDS-SR$_{30}$ total score and the HDRS were all significant and large at baseline and follow-up. Moreover, the correlational analyses also showed a strong relationship between change scores in the IDS-SR$_{30}$ factors and change scores in the HDRS over time, which indicates adequate external longitudinal construct validity.
(2) We have compared the sensitivity to change of the IDS-SR\textsubscript{30} and the HDRS. Please, find below the modifications made in the manuscript:

**Page 9. Differences in sensitivity to change.** We compared the sensitivity to change of the IDS-SR\textsubscript{30} and the HDRS over the 10±2 weeks period by means of a 2x2 repeated measures ANOVA with time (baseline and follow-up) and instrument (IDS-SR\textsubscript{30} and HDRS) as factors.

**Page 12. Differences in sensitivity to change.**
The 2x2 ANOVA yielded a significant main effect for time, $F(1, 3187)= 5655.33$, $p< 0.001$, but not for instrument $F(1, 3187)= 1.37$, $p= 0.24$. The interaction effect did not reach statistical significance, $F(1, 3187)= 3.084$, $p= 0.08$. In other words, both instruments were equally sensitive to change.

**Page 13.** This indicates a good convergent validity. The severity of depressive symptoms in our sample decreased across time regardless of the instrument, which means that the IDS-SR\textsubscript{30} and the HDRS were equally sensitive to change.

(3) In our opinion the cut-off values of the HDRS and the IDS-SR\textsubscript{30} have been sufficiently studied in the literature. We mention in the Introduction of the manuscript that “In the self-rated version (IDS-SR) a cut-off-point of 18 or above indicates the presence of clinical relevant depressive symptomatology [14]”.

Minor essential Revisions

1. **Information is lacking on how these patients were diagnosed prior to inclusion in the study. Was a diagnostic interview part of the intake procedure? Was diagnosis based on clinical impression of the psychiatrist? The authors should offer some clarification.**

As we state in the study design section, to meet DSM-IV criteria for depression was an inclusion criteria. As can be seen in the instrument section we have added a sentence explaining that the DSM-IV criteria for major depressive episode were included in the case report form that psychiatrist used in the first assessment of the patient.

2. **The HDRS is a rating scale and was probably completed by the psychiatrist. Was any training with the instrument provided and/or required?**

The HDRS was completed by the psychiatrist. It was selected because of the number of psychiatrist that had to include patients. There was no training with the instrument as this scale has been widely used by most psychiatrists in clinical practice and become a standard in clinical trials.
3. The authors mention some adaptations to the South American version of the IDS. To what extent? Were they far-reaching? Is this the first evaluation of the Spanish version and, if not, how do the findings compare to previous evaluations of Spanish (or South American) versions.

In the IDS official website there are some South American translations of the different versions of the IDS. However, as far as we know, no validation has been published. That is what the authors told us when we ask for permission. Although we do not know, probably southamerican versions have been used only in clinical practice and not in epidemiological published studies.

4. Page 12 top: items 9 and 24 were assigned to the anxiety/somatic subscale, 30 tot the Mood/Cognition subscale.

We have corrected the mistake.

5. In the discussion, the authors mention that excluding item 4 may improve the internal consistency. It does or it doesn’t. The present data do not favour the position that it does, as the internal consistency only improves marginally according to the baseline data and not in the follow-up data.

As correctly the rewier suggest this sentence was not correct and we have eliminated it.

Discretionary Revisions

1. Self-report measures can be used to screen for the presence of a disorder or to assess the severity of a condition, but not to diagnose. As such the IDS is not a diagnostic tool, but rather an assessment instrument.

We find the comment really adequate so we have modified the termes in the introduction.

2. The manuscript could benefit from a thorough check by a native speaker as grammatical errors abound throughout the text. Furthermore, the text could be formulated more succinctly. An example: “Among the strengths of the present study it must be highlighted the large representative sample included, which avoids patient bias and makes the results...” could read: “A strength of the present study is the large and representative sample which makes the results...”. Another one: “The main limitation of this study is that we did not assess the questionnaire in primary care.” should read: “… that we did not administer the questionnaire to primary care patients.”

We have corrected the sentences that the reviewer says and a native speaker has checked and corrected the grammatical errors.
Third review

2. Are the methods appropriate and well described? Generally yes. However, I think the authors need to pay a bit more attention to the lack of parsimony produced by adding an additional factor (the author of the scale agrees that it is not unidimensional). The differences in the relevant descriptive statistics (TLI, CFI, and RMSEA) are fairly small. It would help if the authors would report a statistic like the AIC which allows non-nested comparisons to be made (it is provided as MPlus output). The cons as well as the pros of the third factor are a proper topic for discussion.

The issue addressed in this comment is the same as the one suggested by reviewer 1 and it has been answered yet.

Finally, we have cited the RESIST study in our revised manuscript as this has now been published.