Author’s response to reviews

Title: Exploration of the psychometric properties of the Clinical Outcomes in Routine Evaluation-Outcome Measure in Ecuador

Authors:

Clara Paz (clara.paz@udla.edu.ec)

Guido Mascialino (guido.mascialino@udla.edu.ec)

Chris Evans (chris@psyctc.org)

Version: 1 Date: 07 Jul 2020

Author’s response to reviews:

Dear Dr. Darren Byrne,

In the last months we were very concerned about what was happening with our paper. For us this is a very important paper, because it presents the psychometric evaluation of a measure that is now in increasing use in Latin America to track psychological distress in clinical and non-clinical populations and particularly to monitor changes and outcomes for psychological intervention. We recently published a scoping review (Paz, C., Mascialino, G., & Evans, C. (2020). Psychological intervention change measurement in Latin America: Where from? Where to? Psychotherapy Research, 0(0), 1–10. https://doi.org/10.1080/10503307.2020.1750728), submitted to that journal after our paper was submitted to BMC Psychology. The review confirms the lack of any measure for these purposes with any substantial psychometric exploration in Latin America in local languages and samples. We believe this underlines the need of this kind of study for the region and believe you will understand that we think that is urgent to be able to offer the Latin American community an initial psychometric exploration of a suitable measure.

At your invitation we have read and addressed the comments of the reviewers and we are submitting a new version of the paper with our response to each of the reviewer’s comment as follows:

Reviewer 1
Comment 1: I don't agree with the authors decision to exclude an Exploratory Factor Analysis (EFA) based on past findings of scale complexity. In my opinion knowledgeable readers will expect to see some type of structural analysis (EFA, CFA or SEM) in a paper of this type. The logic for excluding the analysis also seems weak, "in light of prior research indicating the CORE-OM has a complex factor structure" (page 7). By the authors own assertion this is the first effort to study the CORE-OM in a Spanish speaking population and culture, so we don't know how past findings may relate to those of the present study. Also leaving the EFA out seems a little bit like cheery picking results (as the structure may not look clean), an appearance that should be avoided in the current climate of transparency. Regardless, I would argue for including an EFA and if the findings turnout to be too complex then the authors can cite the past findings (Trujillo et al.,) to show consistency.

Response: We are puzzled by this, there is no EFA in Trujillo et al. (2016) for precisely the reasons there is no such analysis in our submission. We understand the reviewer’s view that readers will expect to see such an analysis but disagree that this is desirable. The arguments about the disadvantages of pursuing factor structures in general mental health and well-being measures are developing but stand against a very powerful orthodoxy as both reviewers’ comments show. The counter-argument was given in the paper as submitted but we have expanded it a little in the methods section. Expanding here, without developing into a paper length review: we believe there simply is not a clean population factor structure of mental health, well-being or psychological distress in any country or culture except when we focus on very specific subsets of problems such anxiety or depression and even there findings vary markedly (e.g. Giusti, E. M., Jonkman, A., Manzoni, G. M., Castelnuovo, G., Terwee, C. B., Roorda, L. D., & Chiarotto, A. (2019). Proposal for improvement of the hospital anxiety and depression scale for the assessment of emotional distress in patients with chronic musculoskeletal pain: A bifactor and item response theory analysis. The Journal of Pain, in press. https://doi.org/10.1016/j.jpain.2019.08.003). This is true for modern short, broad coverage measures such as the CORE-OM and the OQ-45 but has clearly been true, historically, for the various versions of the GHQ, SCL and BSI. Not only is there not a single clean factorial structure cross-sectionally but having such a structure is not the logic behind these measures at least when used as measures of change as they increasingly are. Change tends to be highly individual with the items touching on key problems for any one client changing more than the other items in a measure meaning that even if there were a neat factor cross-sectional factor structure of such measures, this would not extend to show measurement invariance on repeat measurements during therapy. While this acceptance of the absence of any clean structure is out of line with conventional psychometric analyses which largely come from trait and ability measurement, the work of Mavranezouli et al. on the CORE-OM has shown how this minimally structured multidimensionality supports the potential translation of scores on such complex measures for health economic evaluation (references 29 and 30 in the paper).
We are most reluctant to show yet another complex factor structure (it is complex and essentially similar to that for the English CORE-OM in the UK and other factor analyses of other translations as noted in the citations in the paper) as we believe this perpetuates a largely totemic use of factor analysis. Between us we have work in progress expanding these arguments theoretically and exploring them empirically both with simulations and in substantive data, some coming toward submission, but we believe the arguments are already persuasive. We have tried, without adding too many words and references, to make the argument clear in the methods section and accept that it was not made in the first submission. Here is the proposed addition.

“No internal structure analyses were conducted (in line with Trujillo et al. [20]). This respects the arguments in Evans [28] that the CORE-OM was never intended to have a domain based factor structure but to have wide coverage of many issues, covering the four domains of well-being, problems, functioning and risk which are complexly interrelated both across individuals and in patterns of change within individuals in therapy. This is congruent with the findings of the expected complexity of structure in Lyne et al. [29] and the detailed work of Mavranezouli et al. [30, 31] showing how the complex structure supports health economic evaluation underlining that conventional cross-sectional psychometric structure neatness can be a disadvantage for short, broad coverage measures designed for evaluation of change as well as state at single time points.”

Comment 2: The concurrent validity instruments used in the study, the SOS-10 & OQ45.2, were also developed and validated in the US with Spanish speaking samples. This is one of the first tests of these scales in a Spanish speaking population and culture, as such it would be of benefit for the authors to also provide basic psychometric data on these scales. I would suggest providing internal consistency along with the scale means and standard deviations for both the SOS-10 & OQ45.2. These data will serve two important purposes; first it will allow the reader to assess the performance of these scales with respect to offering concurrent validity and second presenting these data may encourage other researchers to evaluate these scales in similar populations thereby increasing the number of scales available for cross-cultural outcomes research.

Response: We are happy to do this, we have included internal consistency, mean and SD for SOS-10 and OQ45 in table 4.
Comment 3: In the discussion section the authors raise a concern (or limitation) regarding the fact that the education level of their sample (65.4% have 12+ years of education) exceeds the national average for the country. In fact, the high education level of the sample may shift it away from representing a Spanish speaking culture and towards Western culture. A similar concern has recently been raised regarding cross cultural research exploring the structure of normal personality. So, I think this is a significant problem and I would encourage the authors to address it as best they can with these data. For example, 23.4% of the study sample has an education level of ≤12 years (that's about 180 subjects). It would seem possible to split the sample into two groups (Education ≤12 & >12) and explore the impact of "education" on test functioning and performance. Also providing the means and standard deviations for these two subgroups could enhance score interpretation.

Response: We have considered this as a post hoc analysis, and we include a paragraph in the results section. No significant differences were found between the two groups:

“Although these were not a priori planned analyses as requested by an anonymous reviewer we explored the differences of the scores with regard to the level of education (12 or less years of education vs. more than 12 years of education) for the community sample. In total 81 participants reported 12 or less years of education, while 227 reported more than 12 years of education. No significant differences in scores for the domain scores (95%CI for the mean difference of Well-being [-.07, .23], Problems [-.13, .16], Functioning [-.01, .27], and Risk [-.11, .03]), non-risk (-.06, .18) or total scores (-.05, .16) were identified between both groups.”

Also, we have included an interpretation of these results in the discussion section:

“However, when comparing the participants with more than 12 years of education, and those with 12 or less years of education, no significant differences were found in the scores. This would suggest that years of education is not associated with response to CORE-OM, though clearly more studies, and accumulation of a larger sample of persons with less than 12 years of education is needed to gain more precision for this finding.”

Reviewer 2
Comment 1: The target user of the study instrument is not clear at all. I think the author need to explain it and provide more examples.

Response: The comment was helpful in order to show the different uses of the measures. In the previous version we have explained the use of the CORE-OM in research settings, but this measure is also extensively used by clinicians. Then we have added information about possible uses of the measure by clinicians and mental health services as follows:

“As intended by the designers, the CORE-OM has been utilized in research for many purposes, such as determining the level of psychological well-being in a given population [8], evaluating the effect of psychological interventions [9, 10], exploring psychotherapy process [11], and as an outcome measurement in randomized controlled trials [12]. The CORE-OM has been also used to generate practice-based evidence, a paradigm that looks for complementing evidence-based practice through the provision of information recovered for practitioners everyday practice [9]. The measure was made available free of reproduction costs to support its use whether in large mental health services, but also to ensure that it could be used in small services or private practice. The information collected using CORE-OM by practitioners can serve for many purposes, one of them as a feedback system of the progress of their clients [13].”

Comment 2: Language editing is needed throughout the manuscript.

Response: We were not struck that the English was unusual. English is the second language of the first author. The second author is effectively bilingual in Spanish and American English and the third author is a unilingual native English speaker. We have been through the paper scrupulously again and found, to our eyes, only very few and small language issues. We believe that the resubmission is in native English, perhaps mid-Atlantic style but North American spelling. If the journal does not agree we are happy to have the issues clarified.

Comment 3: People receiving psychotherapy o psychotropic medication were excluded. Please provide justification
Response: We have clarified the objective of the study, which is to explore the psychometrics properties of the CORE-OM in a non-clinical population, and that is the reason why possible clinical population was excluded, as we expected that population will present higher levels of distress. The next study is to use the CORE-OM in a clinical sample, but for that we needed a non-clinical sample for comparison. The re-worded paragraph in be found in the methods section in procedures:

“This is a psychometric exploratory study that aims to evaluate the psychometric properties of the Spanish translation of the CORE-OM [20] in Ecuador in a non-clinical sample. Data collection occurred from December 2017 to May 2018. Participants were excluded if they reported receiving psychotherapy treatment and/or if they were taking psychotropic medication. They were excluded as they can be considered a clinical population and the immediate focus of the study pending the accumulation of a clinical sample was on the properties of the measure in the non-clinical population.”

Comment 4: The sample size justification is a bit confusing. Please revise it

Response: We have changed it to this:

“The total sample consisted of two subgroups: a student subsample and a community subsample as in previous studies [15, 20]. Having those two sub-samples allow the comparison of the properties with those studies as, though easier to recruit, student samples are clearly not representative of the entire non-clinical population. Convenience sampling was used given low funding and the exploratory nature of the study. The sample size calculation took into consideration the various analyses planned (of acceptability, internal reliability, convergent correlations, age and gender effects and test-retest stability in the student subsample). Minimum sample sizes to give good power to detect meaningful differences from the Spanish [20] findings varied across those analyses. The key analysis with the lowest power was comparison of completion rates for acceptability. This would have a 95% confidence interval from .93 to .96 around an observed completion rate of .95 for a sample of 700, which seemed sufficiently precise for comparison with existing findings. For the test-retest study a sample of 100 would give power to detect the key time 2 to time 3 mean change, which was expected to be small. Data collection was planned to continue until the study period was completed with the period based on resources but which, allowing for a high estimated refusal rate would we thought guarantee the minimum sample sizes being exceeded. In the event better than expected recruitment this resulted in larger than minimal samples.”
Comment 5: Why were there two groups of sample (students vs community)

Response: We considered those two samples in order to replicate and compare our data with previous studies. Probably, it was not well explained in the previous version, but we have added the explanation:

“The total sample consisted of two subgroups: a student subsample and a community subsample as in previous studies [15, 20]. Having those two sub-samples allow the comparison of the properties with those studies as, though easier to recruit, student samples are clearly not representative of the entire non-clinical population.”

Comment 6: I suggest you should have a brief literature review about the psychometric properties of the study Instrument in the introduction section to show that this measurement is robust in other populations.

Response: We have moved this information from the Methods section to the introduction and made some changes in wording for better understanding.

“Whilst not expected to transfer without any changes of meaning, of psychometrics or of referential score distributions across all cultures and languages, the CORE-OM was hoped to transfer across many and it has now been translated into over 25 languages [14]. All these translations have followed the CST protocol [14] and respected the philosophy to offer translations which might be acceptable to very diverse patients/clients. Psychometric properties in the original UK exploration were good [15]. Internal consistency ranged from acceptable to excellent (α = .75 to .94), test-retest reliability was excellent (ρ = .91), and convergent validity was good as evidenced by strong correlations with the Beck’s Depression Inventory-II [16] (r = .85) and the Symptom Checklist 90-Revised [17] (r = .88). Similar explorations in Portuguese[18] and Icelandic[19] versions have shown comparable psychometric properties to the original UK English version[5]. The psychometric properties of the Spanish version [20] were also good. Analysis revealed acceptable to excellent internal consistency (α range = .73 - .94), adequate to good test-retest reliability (ρ = .76 - .87) except for the Risk domain (ρ = .45), and good convergent validity with the Beck’s Depression Inventory-II [16] (r = .83) and the Symptom Checklist 90-Revised [17] (r = .79).”

Comment 7: I do not understand why factor analysis was not conducted
Response: Our arguments against factor analysis of measures like the CORE-OM are given above in the response of comment 1 from reviewer 1.

Comment 8: Intraclass correlation coefficient is the gold standard for test-retest reliability. The use of Spearman's rank correlation coefficient is wrong

Response: We are a little surprised the reviewer is so trenchant about this. Clearly there are a number of different ways to measure/parametrize test-retest reliability. We chose the Spearman correlation for consistency with the original 2002 UK report and the Trujillo et al. 2016 report on the Spanish data as it is the only way to preserve comparability with those two reports. Perhaps ironically the Spearman was forced on the 2002 paper by a peer reviewer who insisted that the Pearson was inappropriate as the data were not Gaussian in distribution: a good example of homage to p values when they are completely irrelevant. As there is no marked non-linearity in the relationship between the scores across occasions, the Spearman correlation is not misleading and is close to the Pearson.

There is a debate about ICCs, certainly they are hugely powerful for inter-rater reliability work and unless fixed rater effects are to be dissected one of the ICCs is almost certainly the best way to summarize that form of reliability. However, technically the intra class correlation is designed for the situation in which observations within the class (hence “intra class”) are interchangeable, as they generally are in inter-rater reliability exploration. However, of course ratings are ordered and not interchangeable in a test-retest study.

Putting that issue aside, there is problem with ICCs that there are multiple ICCs and too many papers do not clarify which they used. Presumably the reviewer is advocating the two-way, single rating model, and the agreement model not the consistency model. The reason we do not see that as “gold standard” is that it penalizes simultaneously for mean shift and for imperfect correlation which seems less informative, and to some extent to mix unreliability (correlation) with invalidity (mean shift). This is not a completely academic issue given the widespread finding of test-retest effects on repeating MH measures in non-clinical samples (see the useful review: Durham CJ, McGrath LD, Burlingame GM, Schaalje GB, Lambert MJ, Davies DR. The Effects of Repeated Administrations on Self-Report and Parent-Report Scales. J Psychoeduc Assess 2002; 20: 240–257 cited in the paper). We prefer to separate the loss of test-retest reliability from imperfect correlation from the mean shift (technically a failure of validity, not reliability). However, we have now reported the ICC in table 2 but retained the Spearman correlation for comparability with previous reports.
Comment 9: You need a table to provide the descriptive statistics of your study sample

Response: We have included the description of the sample in the paragraph, we already have 6 tables and we consider that all the information is covered in the first and second paragraph of the Results section.

“Of the 1,061 persons invited to participate, 587 were female (55.3%); gender was missing for three (0.3%). After refusals and exclusions, the sample consisted of 886 persons (Figure 1), 479 of whom were female. The slight excess of females excluded/refusing was not statistically significant (X^2 (1) = 3.002, p = .08). The female to male ratio was higher in the student sample (58.6%) than in the community sample (47.5%; X^2 (1) = 9.84, p = .002). The ages ranged from 18 to 79 (M = 28.99, SD = 11.89). In the community, 1.4% of the participants reported completing elementary school (6 years), 22.0% had completed high school (12 years), 65.4% reported more than twelve years of education, and for 11.2% this data was missing. Of the student sample, 49.0% were psychology majors, 50.8% non-psychology majors, and for 0.2% this information was missing.”

Also, in table 5 all the referential data is presented for each sample split by gender.

We hope that this resubmission addresses the concerns of the reviewers and we note that first reviewer conclusion was very supportive on the paper:

“Still overall, this is a well-done translation study of a widely used outcome measure, CORE-OM, and should add meaningfully to the current literature and research practice.”

We look forward for your decision.

Kind regards,

Clara Paz, Guido Mascialino, & Chris Evans.