Author’s response to reviews

Title: A revised short version of the Compassionate Love Scale for Humanity (CLS-H-SF): Evidence from Item Response Theory Analyses and Validity Testing

Authors:
FRANCESCA CHIESI (francesca.chiesi@unifi.it)
Chloe Lau (clau263@uwo.ca)
Donald Saklofske (don.saklofske@uwo.ca)

Version: 2 Date: 07 Jan 2020

Author’s response to reviews:

Dear Dr. Balsamo,

We were pleased to hear that our manuscript "A revised short version of the Compassionate Love Scale for Humanity (CLS-H-SF): Evidence from Item Response Theory Analyses and Validity Testing" (PSYO-D-19-00146R1) is being considered for publication in BMC Psychology, subject to a revision and response to the comments raised by the reviewers.

We thank the reviewers for the helpful comments and suggestions and have modified the manuscript accordingly. We enclosed a copy of the revised manuscript as well as our point-by-point responses to the comments made by the reviewers. In the revised manuscript, we highlighted all changes made to the earlier version.

We would like to take this opportunity to express our sincere gratitude for both the reviews and the opportunity to submit a revised copy of the manuscript. We do thank you for helping us to improve the quality of our manuscript.

Kind regards,

Corresponding Author of the Manuscript

Reviewer #1:

Comment: 1. First of all, the primary purpose of this study, shortening the 21-item, psychometrically sound CLS-H, needs stronger justifications in the article. The scale already has a reasonable number of items and shouldn't be a burden to most of its potential respondents. Is it of any major importance to shorten it? The authors do discuss the weaknesses and limitations of
Santa Clara brief compassion scale (SCBCS) along the way. But the main question should not be "what's wrong with SCBCS," but indeed "what's wrong with CLS-H?" Moreover, wouldn't a shortened version impair the "satisfactory content validity" (p. 4, L. 80) of the scale?

Response: We thank the reviewer for taking the time to read the manuscript and for providing sincere suggestions to help us improve the quality of our manuscript. We hope to have addressed your concerns in our modifications to your suggestions.

Specifically, taking into account the Reviewer’s concern, we tried to justify better the need to shorten the CLS-H stressing two points:

a. Whereas the scale already has a reasonable number of items and shouldn't be a burden to most of its potential respondents, when developing a survey or a protocol to assess compassionate love and many other variables of interest, even small reduction can be useful. For example, when proposing a survey, it is important to reduce the burden as much as possible for the respondent (e.g., healthcare workers, social workers) to avoid boredom during administration and to prevent they perceive the questionnaires as time-demanding and redundant. To give some examples in line with this consideration, recently Xia et al. (2019) shortened a scale from 23 to 12 items, and Morean et al. (2014) proposed a brief version consisting of 7 items derived from a 13-item scale. Brevity would also allow researchers and clinicians to administer a larger number of measures for specific screeners.

b. Some items of the original form are not specific to compassionate love but refer to different constructs, such as empathy, kindness, and altruism. With a short form, we used a data-driven approach as we wanted to retain the greatest amount of information for the unidimensional trait of compassionate love. We excluded several items that had a larger amount of measurement noise (i.e., low discrimination values) using the IRT-driven selection procedure. As such, the content validity of the current shortened scale was not impaired, but is some way, it increases. This can be found on page 5 where we further elaborated on these points.

Finally, we agree that the main point is not the weakness of the SCBCS that was introduced to justify the need for a different and stronger short version.

Comment: In addition, I kind of see how giving a CLS-H "in healthcare, workplace, educational settings" (line 100) could take place, but I am not fully convinced. I would have liked to see a stronger justification for the usefulness of the scale. Administer it in educational/ workplace settings to what end? What would the employers do with this information???

Response: We would like to thank the reviewer for the opportunity to explain better this point. To clarify this point we added some references to studies showing that compassion is believed to have a wide range of benefits in healthcare, including better/higher clinical outcomes, quality of care, and satisfaction with services, individual well-being and mental health. Similarly, some studies suggested that compassion fosters school teachers’ engagement and subjective well-being
Comment: 2. similarly, there's no theoretical justification for investigating the Differential Item Functioning across gender groups. While it seems intuitive that males and females may, on average, demonstrate differing levels of compassionate love, what makes it plausible that they interact with the CLS-H items differently?

Response: We thank the reviewer for this suggestion. Whereas, we are convinced that measurement invariance is an added value of a scale, we agree that this analysis might appear partly inconsistent inside the manuscript. Additionally, following the Reviewer’s methodological concerns about DIF (see below), we performed further analyses and we concluded that item 12 does not actually show DIF. Since the reason to perform this analysis was the item selection (for example, similarly to Teresi et al., 2012), we agree that DIF analysis results are poorly justified. For this reason, we have now decided to delete this part thanks to the suggestion of the reviewer.

Comment: 3. A typo on p. 8, L. 202: it should be "NNFI and CFI &gt; .95," not "&lt;="

Response: We thank the reviewer for noticing this typo. It was corrected accordingly and this correction can be found on page 9.

Comment: 4. Methodological concerns:

4.1 To the best of my knowledge, IRT is barely used for the purpose of shortening a scale (though Rasch or other IRT models are often used in scale development). This is because the IRT parameters are estimated with the entire data set as input data; therefore, removing any one item would result in changes in the item parameters (even though IRT is "sample-independent").

Thus, the "'better than average information' criterion" (p. 9. L. 213) is post hoc, arbitrary, and METHODOLOGICALLY VERY CONCERNING. I also read the cited source for this criterion (Huang et al., 2017), only to find that Huang et al. also did it arbitrarily, without citing any credible source to endorse this important methodological decision.

Response: Concerning the first part of the comment, we beg to disagree regarding the utility of IRT in shortening a scale. Indeed, there are several works that use (to give just some examples, or van Damme et al., 2010; Kolva et al. 2017; Xia et al., 2019) or explain how to use (Edelen &amp; Reeve, 2007) IRT to shortening a scale. Items were chosen based on evaluation of each individual item information functions (IIF) graphically represented by the item information curve (IIC), in which the area above the IICs were individually examined (Bortolotti, Tezza, de Andrade, Bornia, &amp; de Sousa Júnior, 2013; Reise &amp; Waller, 2009). Through applying IRT, items that maximize measurement precision across different ranges of the latent trait continuum may be selected for an effective, precise, and non-redundant short-form.
We acknowledge that the "better than average information criterion" might appear arbitrary and that the source does not allow endorsing this methodological decision. Thus, to guide item selection, we examined the item information functions, without any specific cut-off criterion, in line with, for example, Edelen & Reeve (2007) and van Damme et al. (2010), which are deemed credible sources.

Comment: 4.2 More importantly, item information function (IIF) is mostly proportional to the discrimination (a) parameter but less related to the threshold (b) parameters. As we may see from Table 1, the items with above average information are also items with the highest discrimination (a) values, while some of the selected items do not seem to cover a reasonably wide range of the latent trait being measured (e.g., b6 for Item 3 is barely 1.00, meaning that this item is incapable of capturing the differences among individuals high in compassionate love)

Response: Thank you for the insightful comment. We agree some items do not seem to cover a reasonably wide range of the measured latent trait, which means that these items are incapable of capturing the differences among individuals very high in compassionate love. Nonetheless, this is a common characteristic of the CLS-H items. These results are consistent with other studies that showed measurement of positive psychological characteristics (e.g., optimism) tend to be more precise at lower ends of the latent trait spectrum (e.g., Chiesi, Galli, PrimI, Innocenti Borgi, & Bonacchi, 2013). Indeed, maximum values for b6 are around 2 with the exception of item 14 and 20. Unfortunately, these two items cover all the range of the latent trait, but conveyed a minimum amount of information (see Figure 1). Items that have a low a value also tend to measure a larger range of the latent trait more precisely. As such, we also observe that the original scale is less precise for the higher levels of the trait (see Figure 2). Thus, looking at the IIFs we selected the item that performed better along a wide range of the trait, being aware of this scale limitation and that none of them covers adequately the higher levels.

Comment: 4.3 For DIF across gender, more detailed information should be given in regard to the selection of anchor items in the several rounds of DIF analysis. Also, should the alpha level be adjusted/corrected to control familywise Type I error rate? Even Item 12 (p = .008) perhaps shouldn't be flagged as demonstrating significant DIF if we consider this. More importantly, what is the theoretical reasons behind the DIF across gender groups? Such should be discussed to provide implications and guidance for future compassion love research. This is only briefly mentioned in the Discussion section with unclear language (i.e., "Additionally, investigating gender differential item functioning . . . that appeared to be biased against gender . . ." should be phrased as "biased against 'males.'")

Response: We thank the reviewer for the insightful comment. Accordingly, we deeply investigated DIF using measure of effect size and actually none of the nine items displayed DIF. Thus, as stated before, we preferred to delete this part that might poorly justify and not be useful for the item selection.
Comment: 4.4 For reliability estimates, the information presented in the article seems a bit confusing. The authors mention Cronbach's alpha (p. 13, L. 288) but is it the CTT-based alpha or a different reliability estimate?

Cronbach's alpha is total, raw score based, and I cannot see how it is estimated at different theta levels (or maybe it's a method I don't know). Also, it seems meaningless to compare the alphas of the original and shortened versions because the alpha coefficient also acts as a function of the item number k. I would like more information about the IRT/theta-based reliability estimates.

Response: Thank you to the reviewer for this comment. We agree that introducing Alpha values might be confounding. Thus, we decided to delete this part and left only the IRT/theta-based reliability estimates. The information (I) can be transformed in $r = I / (1+I)$, which is similar to MacDonald’s $\pi$ (MacDonald, 2013). Otherwise, we can use the following formula: $1 - (1/I)$ (Thissen, 2000), which is more often employed when using IRT. Thus, we used the latter transformation and we added the reference.

Comment: 4.5 Figure 1 should be re-created to reflect the loss of information of the shortened version (vertical axes should be aligned; or consider combining the TIFs in one graph).

Response: Thank you for the reviewer for catching this. The figure was modified and the axes of the figures are now the same.

Comment: 5. Some minor writing/stylistic issues:

5.1 I suggest verification that the abbreviated journal names are what's required. They impede the ability to locate the article (which is the main purpose of the references section after all).

Response: We checked the abbreviations and now they should be correct.

5.2 One stylistic comment: the authors use the expression "in order to" very many times throughout the article. I suggest revising all instances into simply "to."

Response: All the “in order” were deleted.

5.3 Some sentences exhibited linguistic quirks/typos (for example, line 90: "this short scale was obtained adopting poor item selection criteria" should probably read: "this short scale was created BY adopting…"). I suggest another round of edits for clarity.

Response: The paper was edited to avoid these kinds of errors.
5.4 One sentence was really hard to follow and requires revision (line 65: "These responses from the individual allow one to pay attention to the needs and miseries of others, thus facilitating an individual to support others around them intimately.") I am honestly not sure what this means.

Response: We deleted the sentence.

5.5 Lastly, a minor point, but stood out prominently to me: make sure to have consistent spacing (one space before and one after) whenever "=" is used throughout the paper, especially in the Measures section.

Response: We checked and amended all the missing spacing.

Reviewer #2:

Comment: According to "Standards for Educational and Psychological Testing" (AERA, APA & NCME, 2014), reliability and validity are not properties of an instrument, and validation is an ongoing process that involves accumulating all relevant evidence to support validity. Strictly speaking, it is less appropriate to express reliability and validity in the sentences like "the reliability and validity of a psychometrically sound short version of the CLS-H" and "validity of both the original and shortened form of the CLS-H". Additionally, the authors indicated "to confirm the short scale internal and external validity in comparison to the original form". Clarifications are needed in terms of what particular internal and external validity evidence this study sought to focus on.

Response: We thank the reviewer for taking the time to read the manuscript and for providing sincere suggestions to help us improve the quality of our manuscript. We hope to have addressed your concerns in our modifications to your suggestions.

Specifically, the mentioned sentences were modified to make them clearer or not misleading. As for validity, the factor structure was intended as internal validity, while external validity was about relationships with external constructs (such as anxiety or self-esteem). To avoid imprecise terms, we used the more general terms validity and factor structure.

Comment: Under the "Methods" section, the participants are reported as "undergraduate students (N=790; 65.8% females)" with ages "ranged from 16 to 36 years (M = 18.93, SD =1.86)." What were the sample distribution and sample characteristics? Was the sample skewed? Would the sample skewness affect the generalizability of the findings of this study?

Response: Thank you for the insightful comment. As the reviewer stated, the age distribution was skewed. Given that we have collected this data from undergraduate participants, we expect generalizability (e.g., different age range, education level, SES) to be a limitation in this study. We acknowledged this point in the limitations on page 9.
Comment: In the dimensionality check for the IRT analysis (p.8), the NNFI and CFI criteria should be \( \geq .95 \), rather than \( \leq .95 \).

Response: We amended the typo.

Comment: In the IRT analysis, given the very low values of b1 for some items as reported in Table 1, it would be helpful to check the frequency of each response category (how many participants endorsed each response category ranging from 1 to 7). This would provide an indication about whether some of the response categories need to be collapsed/combined.

Response: Thank you to the reviewer for the insightful suggestion. We observed that item 18, 19, and 21 have very low frequency rates (<15) of the first response option. Thus, as suggested, we collapsed the option 1 and 2 and we repeated the analysis. Results don’t change substantially, as can be seen in Table 1.

Comment: In the short form construction, the authors claimed to select “the items that offered higher information.” Since item information varies at different levels of the latent trait (theta), more technical details are needed about this item selection process. Further, elaborations are also needed to explain how “the shape of the item information function was also considered when selecting items that had high discrimination parameters.” Was there any particular range of theta that had to be considered? In addition to psychometric criteria, how about the content/construct considerations? Does the short form have a good coverge of all key conceptual elements of the compassion construct?

Response: Thank you to the reviewer for the insightful comment. Since item information varies at different levels of the latent trait (theta), we included the IFFs (Figure 1), to better illustrate the motivations of our choices. Specifically, in general CLS-H items do not seem to cover higher theta levels, which means that these items are incapable of capturing the differences among individuals very high in compassionate love. Only item 14 and 20 cover a wider range but, unfortunately, conveyed a minimum amount of information (see Figure 1). As such, also the original scale is less precise for the higher levels of the trait (see Figure 2). Thus, looking at the IIFs we selected the item that performed better along a wide range of the trait, being aware of this scale limitation and that none of them covers adequately the higher levels. However, the selection was not focused on a particular range of theta since we did not have specific theoretical or methodological consideration to do that (as a cut-off value around which the maximum precision should be obtained).

As for content/construct considerations, we checked the content of the items selected applying the IRT-driven procedure and we noticed they addressed the key features of compassion as being aware of people’s adversity and discomfort (e.g. item 4), responding emotionally to their suffering (e.g. item 10), and desiring to act to alleviate people’s misery and distress (e.g. item 12). Moreover, most of the removed items were not about compassionate love but refer to different constructs, as altruism (e.g., item 11), empathy (e.g., item 18), and kindness (e.g., item...
21). As such, the content validity of the current shortened scale is not impaired, but rather increases.

Comment: In the "reliability" subsection under "Results" (p.12), the term "reliability" is better phrased as "measurement precision" in the IRT context so as to avoid potential ambiguity and distinguish from the reliability coefficient under the CTT framework.

Response: Thank you for the suggestion. We changed the term accordingly.

Comment: In the "validity" subsection under "Results" (Table 3, p.14), were the correlations performed on the basis of theta or observed scores? Clarification is needed here.

Response: Thank you to the reviewer for this comment. Correlations were now performed using the IRT scores computed using the two different set of items (long and short version). We also added, as suggested by Edelen & Reise (2007), the IRT score correlation and comparison.

Comment: Under the "Discussion" section, in addition to the discussions about the psychometric contributions of the present study, it is crucial to discuss/elaborate in-depth on the conceptual/theoretical contributions of this study as well as the values and practical implications of the findings in compassion research and practice.

Response: Thank you to the reviewer for the suggestion. Taking into account this comment, we tried to better explain both in the Introduction and the Discussion, the contributions of this study as well as the practical implications.

First of all, we detailed more the importance of obtaining a well-performing shorter scale in light of the limitations of the original scale and the previously developed short version. Whereas the scale has a reasonable number of items, when developing a survey or a protocol to assess compassionate love along with many other variables (i.e., many other scales) even a small reduction can be useful. For example when proposing a survey to social or healthcare workers, it is important to reduce when possible the burden for the respondent to avoid boredom during administration and to prevent perceiving the questionnaires as time-demanding and redundant. To give some examples in line with this consideration, recently Xia et al. (2019) shortened a scale from 23 to 12 items, and Morean et al. 2014 proposed a brief version consisting of 7 items from a 13-item scale.

Then, we illustrated the various fields of application. To clarify this point we added some references to studies showing that compassion is believed to have a wide range of benefits in healthcare, including better/higher clinical outcomes, quality of care, and satisfaction with services, individual well-being and mental health. Similarly, some studies suggested that compassion fosters school teachers’ engagement and subjective well-being at work and, more in general, affects the workplace climate. Therefore, the current study makes a contribution to the research on compassionate love focusing on the measurement issue. Indeed, employing IRT
analyses techniques we developed a precise and valid short form. This concise and efficient measure can provide an added value in research and practice because it can be used to identify expressions of compassionate love across various contexts, to highlight the positive effects of compassionate love, to implement interventions that seek to enhance people's compassion.

Comment: The manuscript would potentially benefit from language editing.

Response: The paper was edited.