

Revision Comments to the Manuscript Nr. 24284 entitled:

“The Peritraumatic Behavior Questionnaire - Observer Rated (PBQ-OR): Validation of the objective version of a measure for combat-related peritraumatic stress”

by Agorastos A., Angkaw A. C., Johnson H. E., Hansen C. J., Cook C. V. and Baker D. G..

We would like to thank Professor Shekhar, the Editor-in Chief, for considering this manuscript for publication and the Reviewers for their favourable evaluation and their detailed comments/suggestions, which helped to considerably improve the overall quality of the manuscript.

We here provide a detailed point-by-point response to the queries of all Reviewers and have included **yellow highlighting** in order to mark applied changes and **green highlighting** to mark already existing parts referring to comments of the Reviewers in the body of the main manuscript.

Responses to comments of Reviewer 1 (#00784126)

Comment 1: *This is a nice article presenting a useful instrument.*

Response: We would like to thank the Reviewer for this positive judgement on our manuscript.

Responses to comments of Reviewer 2 (#03525921)

Comment 1: *This is an interesting study which attempted to look at measures of peritraumatic stress during combat deployment using the PBQ-OR and compare this with a validated post-deployment self-report questionnaire PBQ-SR in order to validate the peritraumatic observer rated questionnaire.*

Response: We would like to thank the Reviewer for these positive comments.

Comment 2: *The timeframe of evaluations was not clearly conveyed early in the paper.*

Response: The study design is presented in the very first sentence of the methods section (green highlighting, page 6). We additionally moved the section on data collection towards top of the methods section (yellow highlighting, pages 6-7), in order to cover this topic earlier in the manuscript.

Comment 3: *There were references made to baseline measurements which in reality were referring to 30 days into deployment, which in all fairness should not be referred to as baseline. It is likely that not collecting data prior to deployment resulted in missing a change from true baseline stress and the levels of stress during deployment.*

Response: There is actually only one passage, where the word baseline is stated in the manuscript (second sentence, page 7 – green highlighting).

“The instructions for PBQ-OR required each symptom to be rated as present only to the extent it was a clear change from **baseline** behaviour for the rated individual, persisting for “a period of time” after exposure to an identifiable stressor”.

However, the word baseline here does not refer to baseline in terms of pre-deployment, but rather to baseline as to “normal” behaviour during deployment (i.e., before a traumatic event). Since the PBQ-OR is a military-specific instrument for the assessment of peritraumatic reactions in the field of operations, pre-deployment stress level should not be evaluated through this questionnaire.

Comment 4: *There are indeed many difficulties with collecting data in theatre and this study had a high rate of non-response or non-return of PBQ-OR questionnaire from raters. It is entirely possible that these raters were embedded with units which were more stressed than raters who had time to complete and return the questionnaires.*

Response: As stated in the Manuscript (Results, green highlighting, page 8): “29 certified male Corpsmen signed informed consent to be included as raters in the PBQ-OR study. Of those, 7 actually returned PBQ-OR ratings with dropouts occurring for different reasons (e.g., non-embedment with a unit, operational schedule, serious injury, loss of data in battle).”

Thus, it was actually rather operational reasons leading to the high drop-out rate of the raters and not stress-related, although this cannot be entirely excluded.

Comment 5: *The main limitation in this study is the low rate of response. Out of 860, only 62 questionnaires were returned for R3, which is a very low rate of response making it difficult to base any solid conclusions on the results. That being said, this study serves as a basis for a larger study which would serve to validate the PBQ-OR and the limited data analysed did show a trend towards it being a valid instrument at measuring peritraumatic stress.*

Response: We agree with the comments of the reviewer and had therefore already pointed out this limitation right at the beginning of our discussion section (green highlighting, page 10). We have additionally included the following statement as the last sentence of our conclusion section (yellow highlighting, page 11) to address the second comment of the Reviewer:

“The ability of the PBQ-OR to serve as a standard questionnaire with practical applicability and incremental validity in military settings should, however, be prospectively validated through additional, larger-scale studies.”

Comment 6: *Intro: This section states that peritraumatic stress can be a “salient clinical indicator” for PTSD, however in military personnel this stress may have a high sensitivity but not be very specific, as it is difficult to decipher who will go on to meet criteria for PTSD.*

Response: We agree with the Reviewer’s comment and have therefore replaced the word “salient” through “very sensitive”.

Comment 7: *Material and Methods: What was the criteria used for PTSD diagnosis?*

Response: This is stated in the section “Measures” (green highlighting, page 7):

A DSM-IV diagnosis of PTSD was made using the well-established F1/I2 scoring rule (Weathers et al. 1999).

Frequency > 1/Intensity > 2 (F1/I2) rule: This was the original scoring rule proposed by Blake et al. (1990). According to this rule, a PTSD symptom is considered present if the frequency of the corresponding CAPS item is rated as 1 or higher and the intensity is rated as a 2 or higher. This roughly corresponds to Blanchard et al.'s (1995) more inclusive Rule of 3,

the difference being that Blanchard et al. also considered a symptom to be present when the frequency was 2 or higher and the intensity was 1 or higher. That is, they considered a symptom to be present when the severity of the corresponding CAPS item (frequency + intensity) was 3 or higher. The test-retest reliability of PTSD diagnosis based on the CAPS F1/I2 rule is $\kappa = .78$, exactly as high as the test-retest reliability using a CAPS severity rule > 4 and almost as reliable as clinician rating with DSM-IV criteria ($\kappa = .80$) (Weathers & Litz, 1994).

Comment 8: *Material and Methods: Were there any data collected prior to deployment?*

Response: There were no pre-deployment data collected for this study.

Comment 9: *Material and Methods: I question why start rating 30 days into a deployment without a baseline? As 30 days into deployment is likely to have already been a major change from participants' real baseline and is likely to mask the change which we are looking for.*

Response: Here we kindly refer to our answer on Reviewer's comment #3.

Comment 10: *Statistics section: What does "by-symptom" mean? Is this a typo?*

Response: By-symptom correlation means symptom-to-symptom (or item-to item) correlation (i.e., correlation of a PBQ-OR symptom item to the same symptom item of PBQ-SR).

Comment 11: *Statistics section: Which correlation coefficients were used? Please list them here.*

Response: This is already stated in the statistic section (green highlighted, page 8):

"Because PBQ-OR ratings were non-normally distributed, Spearman's rho (ρ) was used to calculate the correlations throughout the analysis."

Comment 12: *Statistics section: Why look for convergence "question by question"? would it not be better to look overall as the questions are not individually validated? A difficulty associated with looking for convergence "question by question" is that you would need to match individual questions. There was no reference made to how this was achieved.*

Response: Internal validity (Cronbach's α), as well as convergent validity (as to post-deployment PBQ-SR, CAPS and PCL) were calculated from the total scores. We do not report internal consistency through item-total correlations (as probably the Reviewer suggests) due to space management issues, since we have already calculated this in our first paper on PBQ-SR.

"The inter-item correlation matrix revealed only positive values (mean .331), indicating proper scoring and that all items measured the same underlying attribute. The corrected item-total correlations varied between .406 and .644 for all items, while deletion of single items did not significantly change the alpha value, suggesting that the sum of the 15 items is acceptable to be used as a measure of peritraumatic reactions in battle-related trauma (Agorastos et al., 2013)."

Since PBQ-SR and PBQ-OR represent two versions of the same questionnaire and the PBQ-SR and PBQ-OR symptom-specific items are identically numbered, we refrained from an additional item-to-total correlation analysis in order to be rather able to focus on the kappa item-to-item analysis.

Comment 13: *Statistics section: I would like to hear more about the construction of the log regression model here and what was included? What symptoms were used?*

Response: Predictive validity of PBQ-OR with respect to PTSD diagnosis at 3 months post-deployment was calculated using logistic regression. This was univariate logistic regression model investigating the association of PBQ-OR total score with PTSD caseness (as described above). No other covariates were included in the model.

Comment 14: *Results section: 860 recruited, R1-248, R2 128, R3 62. Very low response rate!*

Response: Here we kindly refer to our answer to Reviewer's comment #5.

Comment 15: *Results section: What does the term "low response rates mean"? are the authors referring to "non response to these questions or low ratings on these questions? Please clarify.*

Response: We agree, that this statement is confusing and therefore replaced “low response rates” through “low ratings”.

Comment 16: *Psychometric properties: Again I would like to have heard more about what was included in the regression model for “PTSD caseness”.*

Response: Here we kindly refer to our answer to Reviewer’s comment #13.

Comment 17: *Discussion: How was sampling error a limitation here?*

Response: We acknowledge the Reviewer’s comment and have specified the type of errors in our manuscript (page 10, yellow highlighting):

“...sampling (e.g., non-response, sampling frame bias, data access) and non-sampling type of errors (e.g., respondent-, interviewer- or questionnaire-specific errors)”.

Comment 5: *Table 1: Why are questions 5, 14 and 15 highlighted? Was this the case for the version completed by study raters?*

Response: As already stated in the Table legend (green highlighting), these questions showed the higher item-to-item significance of correlation to PBQ-SR items. This is also stated in the results section (page 9, green highlighting):

“... best agreement between the observer and subjective ratings in questions relating to perception of mortal peril (#14) (report order 1: $\rho = .41$, $p < .001$; report order 2: $\rho = .58$, $p < .001$; report order 3: $\rho = .50$, $p < .001$) desire for revenge (#5) (report order 1: $\rho = .38$, $p < .001$; report order 2: $\rho = .44$, $p < .001$; report order 3: $\rho = .61$, $p < .001$), and experience of intense physical reactions to combat (#15) (report order 1: $\rho = .34$, $p < .001$; report order 2: $\rho = .41$, $p < .001$; report order 3: $\rho = .53$, $p < .001$).”

Comment 5: *Table 2: How can the standard deviation be larger than the mean score?*

Response: There is no reason why the standard deviation cannot be higher than the mean score. This simply depends on the N, the distribution and the range of the relative score. In our study, we are dealing with a large N, a skewed distribution and a small scoring in

a broader scoring range, which admits $SD > \text{mean}$ and simply suggests a larger spread-out of the data.