PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Evaluation of psychometric properties of the German Hospital Survey on Patient Safety Culture and its potential for cross-cultural comparisons: A cross-sectional study.
AUTHORS	Gambashidze, Nikoloz; Hammer, Antje; Broesterhaus, Mareen; Manser, Tanja

VERSION 1 - REVIEW

REVIEWER	Phillip Phan, PhD Johns Hopkins University and Medicine United States of America
REVIEW RETURNED	10-Jul-2017

GENERAL COMMENTS	Factor loading of >.40 is a very low number of scale validation. Given that the HSOPC has been used in numerous studies, the factor loading cutoff should be much higher. Set the loading, in series, to .50, .55, .60 and see how many items you lose for each sub-scale. You might get more efficient sub-scales this way. Given that you are validating an existing scale, you should also report predictive validity. The HSOPC includes 2 dependent variables. What is the variance explained by the dimensions on the dependent variables? How does this compare with past studies using HSOPC? Alpha of .60 is ok for exploratory FA. However, for this study because it uses an existing pre-validated instrument, alpha should be set to .70 (Cohen and Cohen). In any case, all of your alphas are <.70, so you should not use the .60 cutoff. Good luck on your revision

REVIEWER	Waterson, Patrick Loughborough University UK
REVIEW RETURNED	10-Jul-2017

GENERAL COMMENTS	I agree with the authors that a strength of this paper is that it represents the first validation of a German version of the HSPSC – I think the authors have done this very well. Where I am a little more sceptical and have more questions is in regard to the second aim as set out on page 6 (use of the instrument's potential for cross-national studies). Here are my questions: - Why do you think only 8 of the 12 dimensions of the HSPSC can be used with confidence in cross–national comparisons?

	- What is it about the other four dimensions, one of which is quite important ‘overall perceptions…’), which makes them vary so much? The reason I’m asking these questions is that all sorts of reasons could be given – type of healthcare system, hospital, sample sizes etc. I’d like to see the discussion being a little more sceptical about how we measure patient safety culture – many other explanations as to how safety culture functions (and what influences it) are possible. I think if I saw a little more of that in the paper then I would be much happier with the conclusions.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

1) Factor loading of >.40 is a very low number of scale validation. Given that the HSOPC has been used in numerous studies, the factor loading cut-off should be much higher. Set the loading, in series, to .50, .55, .60 and see how many items you lose for each sub-scale. You might get more efficient sub-scales this way.

Response: Cut-off of 0.4 was set according to the recommendations from the literature (Fields 2012, Hair 2014). Also, 9 out of 11 models analyzed in the manuscript reported factor loadings >0.4. However, we have conducted the factor analysis according to your advice (>0.5) with the following results:

a. 2 more items, namely B3 and F2 were eliminated from the model.

b. The factor structure remained unchanged.

c. Confirmatory factor analysis resulted in better results: (RMSEA=0.05; SRMR=0.05; GFI=0.92; CFI=0.95; TLI/NNFT=0.94) compared to (RMSEA=0.05; SRMR=0.05; GFI=0.90; CFI=0.93; TLI/NNFT=0.91).

d. In the new model factor 06 (Supervisor/manager expectations/actions) is left with only two items (B1 and B2), which hinders the factor’s internal consistency (Cronbach’s alpha 0.67).

While we agree, that a higher cut-off value (e.g. 0.5) for factor loading does result in a more robust factor structure, in this case the change in factor model does not significantly affect results of the study. Therefore, and in order to have the study results in line with other researchers in the field, we decided to stick to the cut-off value of 0.4.

The following sentence was added to results part of the manuscript to underline the reasons for using this benchmark: “We considered factor loadings ≥0.4 as significant, as this cut-off value was typically used in similar studies [4–6, 10–12, 14–16] and was supported by the literature [22, 23].”

2) Given that you are validating an existing scale, you should also report predictive validity. The HSOPC includes 2 dependent variables. What is the variance explained by the dimensions on the dependent variables? How does this compare with past studies using HSOPC?

Response: Thank you for highlighting this part of the analysis. We added construct validity analysis according to Sorra and Nieva to the corresponding parts of the manuscript:

Methods: “Construct validity. By calculating average of corresponding non-missing items we calculated mean values for each dimension for the original 12-factor model and for the new model that emerged from EFA. Pearson’s correlations were evaluated between dimensions in each model. We expected low to moderate correlations between dimensions. However, correlations >0.85 would indicate possible multicollinearity [4, 22]. We also evaluated the correlations between dimensions of both models with two single item outcome variables – Patient safety grade and Number of incidents reported.”

Results: “Construct validity. Correlation between dimensions of original 12-factor model were between 0.10 and 0.61 (p<0.01). All 12 dimensions were positively correlated with the outcome variable Patient safety grade (correlations between 0.26 and 0.70, p<0.01). Dimensions of 8-factor model from EFA were also positively inter-correlated (0.18-0.54, p<0.01) and positively correlated with the outcome variable Patient safety grade (0.29-0.58, p<0.01). All dimensions in both factor models resulted in no

or week correlation (<0.2) with outcome variable Number of events reported. All correlations are presented in the online Appendix 1."
Appendix 1 was added to the manuscript presenting the table with "Pearson's correlations between single item outcome variables (Patient safety grade (E1) and Number of events reported (G1)) and HSPSC dimensions in two factor models."

3) Alpha of .60 is ok for exploratory FA. However, for this study because it uses an existing pre-validated instrument, alpha should be set to .70 (Cohen and Cohen). In any case, all of your alphas are <.70, so you should not use the .60 cut-off.
Response: We agree, according to the literature, cut-off value for good Cronbach's alpha should be set to ≥0.7.
The following sentence was changed in methods part: "In their exploratory study Sorra and Nieva [4] considered Cronbach's alpha ≥0.6 as acceptable. We used Cronbach's alpha ≥0.7, as it is typically used in later studies using the HSPSC [5, 6, 9, 11, 14, 15, 17, 19], and is well supported by the literature [22, 23]."
Table 5 was updated to reflect acceptable Cronbach's alpha ≥0.7.
In the results section, the paragraph "Internal consistency" was reformulated to reflect acceptable Cronbach's alpha ≥0.7. The revised paragraph reads as follows: "The original 12-factor model demonstrated good Cronbach's alpha for all dimensions except Organizational learning – Continuous improvement (0.68) and Communication openness (0.64). Cronbach's alpha for dimensions of 8-factor model were between 0.73 and 0.87. Two dimensions, Teamwork within units and Communication openness, demonstrated consistently low alphas in other factor models analyzed. Three dimensions, Nonpunitive Response to Error, Staffing and Handoffs & Transitions, had lower than 0.7 values only in one or two of analyzed models. Cronbach's alpha for the remaining seven dimensions in all analyzed models was ≥0.7, if present in the model (table 5)."

Reviewer: 2
I agree with the authors that a strength of this paper is that it represents the first validation of a German version of the HSPSC – I think the authors have done this very well. Where I am a little more sceptical and have more questions is in regard to the second aim as set out on page 6 (use of the instrument's potential for cross-national studies). Here are my questions:

1) Why do you think only 8 of the 12 dimensions of the HSPSC can be used with confidence in cross–national comparisons?
Response: The main reason we think that these dimension can be used in cross-national studies is that they were present in at least 10 out of analysed 12 models. Compositions of these dimensions vary slightly across different models, but this is mostly caused by adding items from other, less stabile dimensions. The composition of patient safety culture (its facets and dimensions) is not yet fully researched. Cross-national research may reveal further characteristics of safety culture that are not being considered in current studies. At this stage, we believe, that the stability of the dimension over different language models is key for first cross-national studies and comparisons. This is not to say, that other dimensions should be excluded but that they have to be interpreted more carefully.
We revised the discussion part in order to make this point clearer (see discussion pp. 20-22).

2) What is it about the other four dimensions, one of which is quite important 'overall perceptions…'), which makes them vary so much?
Response: We agree, that exploring the reasons/mechanisms behind fluctuations in factor structure will allow not only for better comparisons of safety culture in international contexts, but also for improving our understanding of patient safety culture itself. However, explaining the reasons for instability of patient safety culture dimensions as proposed by Sorra and Nieva is far beyond the scope of this paper. Based on our analyses one could speculate that because the dimensions proposed by the authors of the instrument are significantly correlated with one another, which on one

hand indicates that they all measure one common construct – patient safety climate, this may also cause a certain degree of collinearity. While translating/adapting the instrument to a new environment, due to linguistic and cultural differences, the exact meaning and perception of specific items may vary significantly. This can cause the originally mild collinearity between dimensions to result in item exchange, merger or item elimination.
We expanded the discussion part in order to clarify this point (see discussion pp. 20-22).

(Reviewer: 2) The reason I'm asking these questions is that all sorts of reasons could be given – type of healthcare system, hospital, sample sizes etc. I'd like to see the discussion being a little more sceptical about how we measure patient safety culture – many other explanations as to how safety culture functions (and what influences it) are possible. I think if I saw a little more of that in the paper then I would be much happier with the conclusions.
Response: Even though we feel, that analysing the conceptual framework of patient safety culture in different environments is not the scope of this study, we highly appreciate the opportunity to expand the discussion to mention various sample and environmental characteristics that may influence the performance of the instrument and that should be considered in comparative studies. We hope that the revised discussion sufficiently reflects these considerations (see discussion pp. 20-22).

In addition to the reviewers'' comments, we made following changes in the manuscript:
1. Word count is changed – abstract: 255, manuscript: 3266.
2. Abstract, methods: revised sentence: "Psychometric properties (e.g. Model fit, internal consistency, construct validity) of the instrument, and comparison of dimensionality across different language translations."
3. Abstract, results, revised sentence: "…good internal consistency (Cronbach's alpha 0.73-0.87) and construct validity."
4. Abstract, results, the word "dimensions" was added to improve readability of the sentence: "Analysis of the dimensionality compared to models from 10 other language versions revealed eight dimensions with relatively stable composition and appearance across different versions and four dimensions requiring further improvement."
5. Methods, p8, CFA, to support logical numbering of tables in text, following sentence was removed – "Indices evaluated in CFA are presented in table 4."
6. Table 1, Table 2, Table 4 and Table 5, all decimal signs ware changed from comma (,) to dot (.).
7. Table 2 – the name of the column with Standard deviations was changed from "St.d." to "SD5". Explanation was added to the annotation: "5 SD – standard deviation"
8. Table 2 – in annotation "N" was replaced with "n".
9. Table 2 – The question E2 was removed from the table and annotation. Results of E1 were corrected.
10. Table 3 – "E1-E2" was deleted from annotation, as these variables are not presented in the table.
11. Table 4 – relevant references were added to the required criteria for the indices.
12. Results, last paragraph – correction was made: "Similarly, the dimension Hospital handoffs and transitions was merged with the dimension Teamwork across hospital units in 5 models,…" to "…in 4 models,…".
13. Discussion, paragraph 1, in sentence "derived for different language versions of the HSPSC." was corrected to "derived from different language versions of the HSPSC."
14. Discussion, paragraphs 3 and 4 were revised / expanded:

"Our analysis of instrument dimensionality across language versions revealed that whilst some dimensions maintain relative stability of appearance and composition across language versions, others vary significantly. When analyzing 12 different factor models, including the original North American 12-factor model and the 8-factor model resulting from our EFA, we found that items from eight dimensions maintain relative stability in appearance and composition over different cultural

adaptations. These dimensions were Teamwork within units, Nonpunitive response to error, Staffing, Supervisor/manager expectations/actions, Frequency of event reporting, Feedback and communication about error, Hospital management support for patient safety and Teamwork across hospital units. The items from these dimensions seem to maintain their coherence and measure one common factor in different language adaptations and different healthcare systems. In contrast the remaining four dimensions, namely Organizational Learning—Continuous improvement, Overall Perceptions of Safety, Communication Openness and Hospital Handoffs & Transitions, appeared in only ≤60% of analyzed models since corresponding items were either removed, or migrated to or merged with other dimensions. Similarly, Hedskoeld [7] revealed a 9-factor model but argues against removing items and dimensions from the instrument, stating that they can still be used to understand and improve patient safety. Even though these dimensions and corresponding items may be very important in studies of patient safety culture, they need to be refined in order to support their stability over different cultural adaptations.
Evaluation of psychometric properties of a translated version of the instrument is important, as only the results of validated instruments can be properly interpreted and used for comparison in local contexts. A number of studies reported that the original 12-factor model did not fit the data well, and alternative factor models were suggested [6–15]. Variation in the factor structure may be partially attributed to the differences between study samples and study populations. These studies differ by setting, sample size, representation of different professional groups and other characteristics, which can have influence on the performance of the instrument, hence should be considered in analysis. Finally the specific characteristics of study population's culture, as well as of local healthcare system influences how the respondents perceive, understand and respond to each individual item in the questionnaire, ultimately altering the factor structure and interpretation of the results."

The revision was made using the track changes mode.

VERSION 2 – REVIEW

REVIEWER	Phillip Phan, PhD Johns Hopkins University and Medicine USA
REVIEW RETURNED	14-Aug-2017

GENERAL COMMENTS	Thank you for revising your manuscript. You have taken my concerns and addressed them all. The only small concern I have it that you used a mean replacement for the missing values on cases with less than 30% missing responses. I'm always nervous about such manipulations. I understand you do this to preserve power but you should rerun the statistics without doing this (using item elimination rather than case elimination for missing values in the procedure is an acceptable way to save power), and see if your results are similar. If the results don't change very much, its better to not do mean replacement, cleaner. Otherwise, this paper is fine and ready for publication.

REVIEWER	Waterson, Patrick Loughborough Uiversity
REVIEW RETURNED	29-Aug-2017

GENERAL COMMENTS	I am happy that the authors have covered my recommendations for revision.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
1) Thank you for revising your manuscript. You have taken my concerns and addressed them all. The only small concern I have it that you used a mean replacement for the missing values on cases with less than 30% missing responses. I'm always nervous about such manipulations. I understand you do this to preserve power but you should rerun the statistics without doing this (using item elimination rather than case elimination for missing values in the procedure is an acceptable way to save power), and see if your results are similar. If the results don't change very much, its better to not do mean replacement, cleaner. Otherwise, this paper is fine and ready for publication.

Response: Thank you for pointing out this issue. We expanded the relevant part of the results to include additional information regarding the missing values.
During our preliminary analysis, we examined the distribution of missing values. In the sample of n=995, the average number of missing values per participant was 1.07 (95% CI - 0.83-1.30). While missing values per participant on average were not high, n=229 cases (23%) had one or more missing values on HSPSC items. Item elimination (pairwise deletion) was not an option, as it is not used in factor analysis. Hence, using the data without imputation would mean excluding about 23% of the data from the analysis.

We did not use mean replacement for missing values, as this method is relatively crude and may result in reduction of variance. Instead we used multiple imputations with expectation maximization algorithm. This way we were able to represent about 98% of the data in our factor analysis. We excluded only the 21 (2.1%) cases with >30% of missing values on HSPSC items.

We expanded relevant part of the results to include additional clarification regarding the missing data – See Results, Page10.
• Results, page 10: “Out of our sample of n=995, 766 responses (76.98%) had no missing values on HSPSC items. 21 responses (2.1%) contained more than 30% missing values on HSPSC items and were thus not included in the analysis. Remaining missing values were imputed using multiple imputations based on the expectation maximization (EM) algorithm. As a result, n=974 cases were available for further analysis.”

Independent of the reviewer’s comments we corrected the following:
• Results, Table 2. Value for Percent of positive responses for variable E1 was corrected.
• Word count was updated: 3312.

These changes are marked in the manuscript. No other changes were made to the manuscript since the last revision.