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)	Development and Validation of the Oxford Pandemic Attitude Scale-COVID-19 (OPAS-C): An Internet-Based Cross-Sectional Study in the United Kingdom and United States
AUTHORS	Kantor, Jonathan; Kantor, Bella Nichole

VERSION 1 – REVIEW

REVIEWER	Jowsey, Tanisha University of Auckland, Centre for Medical and Health Sciences Education
REVIEW RETURNED	04-Oct-2020

GENERAL COMMENTS	Thank you for undertaking this important work. The study is of utmost importance and the questionnaire is well thought through. The items make sense. The analysis is appropriate and sophisticated, all the right statistical stuff has been done. The paper is excellent. I note a couple of points below that are merely for polishing: p3. you write "online focus groups" - with whom? It would be helpful here to write "online focus groups with anonymous members of the public." or something to that effect. p3. ethics: You write "this work was approved by the AHIRB." I assume you mean the ethical elements of this work? Perhaps this sentence could be reframed so that it makes it clear that you are talking about ethics here? P4. "all respondents were rewarded with a small payment (<$1). I'm not suggesting you need to change anything here. I'm simply marveling at a $1 incentive. Do you think that makes a difference? Fascinating. Perhaps its worth mentioning in your discussion the impact you think this made. I was interested to see the whole thing only takes three minutes to complete - impressive. I would think this is also a strength of your study. P5. It would be good if you could make clear when you are talking about American dollars by using USD$ to indicate. This is an attitudes survey but attitudes are often built on experience and I found myself wondering whether the survey could also collect if participants had actually had covid-19, and whether they had a close family member or friend need intubation or die as a result of covid-19? I would think that this could absolutely inform their attitudes. I might have missed it, but if not, it might be worth a mention in your discussion. Thank you again and kind wishes,

REVIEWER	Moshagen , Morten
	Ulm University
REVIEW RETURNED	23-Oct-2020

| **GENERAL COMMENTS** | The manuscript reports on the development of a scale measuring covid-related responses, such as emotions, attitudes, and beliefs. Overall, I like the paper and think that covid-research might benefit from such a broadband measure. However, I also have some comments that should be addressed in a revision. In particular, a major issue is the severe imbalance of the number of items by scale and, even more importantly, that several of the resulting scales are measured by two items only. The authors should attempt to select items in a way that (a) the number of items by scale does not vary too much and (b) each scale is indicated by at least 3 items. It might also make sense to collapse some dimensions (see my 7th comment), e.g. stress/fear, loneliness/community and maybe NCI/vaccine (as both relate to doing what authorities say).

Specific:

(1) I'd like to read a bit more about the domains, in particular whether domains emerged that were not considered further (and if so, why). Second, the domains are obviously related. For example, stress, fear, and anxiety could easily be merged into a single domain, so I wonder what was the rationale to keep them apart.

(2) How and how many items were generated for each domain?

(3) Though I doubt that results will change, the decision on the number of EFA factors to retain should nevertheless be made in line with recent recommendations, 10.1037/met0000200 . I'd also like to see the largest eigenvalues.

(4) As the authors mention themselves, it does not appear reasonable to drop items based on low interitem correlations, simply because some domains (say fatalism) may be independent of the other dimensions, so we would not expect any meaningful cross-domain correlation.

(5) Convergent/divergent validity was merely assessed through the relations to a single item each, so these results should not be overstated. I recommend something along the lines of "preliminary validity evidence".

(6) p.5: "items were deleted due to low polychoric and polyserial correlations, unifactorial outcomes, duplication, or poor correlations.". The item-reduction process is unclear. What is meant by "poor correlations" and how does that differ from "low...correlations"? What are "unifactorial outcomes"? And why were item-duplicates considered at all?
I suggest to remove items when they exhibited very strong correlations to another item (duplicates) and afterwards just consider the factor loadings and maybe endorsement rates.

(7) Please report the factor intercorrelations.

(8) Just report the robust SEM results. |

(9) I doubt a "total OPAS-C score" makes sense. If there are 7 domains, what should the total score express? Relatedly, alpha for the full scale does not seem to make sense either, so please report reliabilities for each subscale.

(10) Please be more precise concerning the multigroup model. Was it a strong invariance model or just a configural invariance model? In either case, either provide the full sequence of invariance testing or just drop these analyses.

(11) How it comes that there's a loading > 1?

Minor:

- Abstract: "The population" => the sample.
- Table 1: just report two decimals and add factor labels

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Comments to the Author:

Thank you for undertaking this important work. The study is of utmost importance and the questionnaire is well thought through. The items make sense. The analysis is appropriate and sophisticated, all the right statistical stuff has been done. The paper is excellent. I note a couple of points below that are merely for polishing:

Thank you for your very kind comments and for taking the time to perform such a detailed review.

p3. you write "online focus groups" - with whom? It would be helpful here to write "online focus groups with anonymous members of the public." or something to that effect.

Thank you for your kind comments and for raising this important issue. We have revised this sentence as you suggested to read: "After a literature review, online focus groups with members of the public were used to develop a set of possible domains of interest."

p3. ethics: You write "this work was approved by the AHIRB." I assume you mean the ethical elements of this work? Perhaps this sentence could be reframed so that it makes it clear that you are talking about ethics here?

Thank you for your comment. We have reframed the sentence as you suggested so it now reads as follows: "This research was approved by the Ascension Health Institutional Review Board." This is consistent with ICMJE recommendations that read as follows: "The Methods section should include a statement indicating that the research was approved by an independent local, regional or national review body (e.g., ethics committee, institutional review board)." See: http://www.icmje.org/recommendations/browse/manuscript- preparation/preparing-for-submission.html

P4. "all respondents were rewarded with a small payment (<$1). I'm not suggesting you need to change anything here. I'm simply marveling at a $1 incentive. Do you think that makes a difference? Fascinating. Perhaps its worth mentioning in your discussion the impact you think this made.

Thank you for raising this fascinating issue. The impact of small financial incentives on research participation is an interesting area of research and is particularly important to consider in light of the potential for financial incentives to paradoxically disincentivize actions by suggesting that they are undesirable. We used Prolific Academic, a UK-based company that specializes in academic survey research, and they have developed panels of respondents that are incentivized in this fashion (this is similar to other less-academically inclined survey panel companies such as Survey Monkey and Amazon Mechanical Turk). Prolific Academic requires all participants to be compensated with the equivalent of at least ~USD$7 per hour. In many cases, respondents are likely motivated more by an interest in engaging in studies and surveys rather than the financial incentive itself.

I was interested to see the whole thing only takes three minutes to complete - impressive. I would think this is also a strength of your study.

Thank you for the kind comment. We have amended the beginning of the paragraph following this statement to clarify that the 3-minute completion time (feasibility) is a strength of the study by beginning it as follows: "Additional strengths of this validation study include..."

P5. It would be good if you could make clear when you are talking about American dollars by using USD$ to indicate.

Thank you for pointing this out; given that we included both US and UK respondents and that the manuscript is being considered by *BMJ Open*, we have changed USD$ to £.

This is an attitudes survey but attitudes are often built on experience and I found myself wondering whether the survey could also collect if participants had actually had covid-19, and whether they had a close family member or friend need intubation or die as a result of covid-

19? I would think that this could absolutely inform their attitudes. I might have missed it, but if not, it might be worth a mention in your discussion.

Thank you for raising this interesting issue. In some of our earlier work that was used to inform the OPAS-C development (https://doi.org/10.3389/fmed.2020.00384 and https://doi.org/10.3389/fpsyt.2020.569083), we did specifically ask about personal and family experience with actual COVID-19 infection. These elements were intentionally not included in the OPAS-C scale itself since the scale is attempting to capture the attitudes themselves rather than the contributors to those attitudes (experiences, personal beliefs, political worldview, attitude to science, etc.). Moreover, to make the final score on the OPAS-C meaningful, we felt that including a scoring element for something that is out of the control of the respondent (COVID-19 exposure) would potentially bias the results since the goal is to capture attitudes (the responses to experience) rather than the experiences themselves. Thank you again for highlighting a fascinating and important area for future research.

Reviewer: 2

Comments to the Author:

The manuscript reports on the development of a scale measuring covid-related responses, such as emotions, attitudes, and beliefs. Overall, I like the paper and think that covid-research might benefit from such a broadband measure. However, I also have some comments that should be addressed in a revision. In particular, a major issue is the severe imbalance of the number of items by scale and, even more importantly, that several of the resulting scales are measured by two items only. The authors should attempt to select items in a way that (a) the number of items by scale does not vary too much and (b) each scale is indicated by at least 3 items. It might also make sense to collapse some dimensions (see my 7th comment), e.g. stress/fear, loneliness/community and maybe NCI/vaccine (as both relate to doing what authorities say).

Thank you for your kind words and your very important comments regarding domain imbalance. It is wonderful to receive such feedback from an expert in factor analysis. You are absolutely spot-on regarding the variation in item numbers between domains. Here, the disparity in item numbers was driven by clinical considerations regarding the width of different domains, and indeed our original set of items to be considered emphasized items relating to stress given earlier work that suggested the importance of pandemic-related stress's effect on the overall response to COVID-19 (see, for example, some of our earlier work here https://doi.org/10.3389/fmed.2020.00384 and here https://doi.org/10.3389/fpsyt.2020.569083 as well as https://dx.doi.org/10.1016%2Fj.janxdis.2020.102232). Some domains, such as stress, are so broad/multidimensional that a larger number of items are needed to fully capture the range of responses, while others, such as vaccine hesitancy, are more narrow and therefore a smaller number of items were felt to be adequate on a clinical basis. Importantly, the number of items included a priori for consideration was driven by clinical considerations regarding the importance of including them in the final scale. While it is common practice to include a

minimum of 3 items in each domain, this is driven in part by statistical concerns regarding whether 2 items would yield convergent solutions in confirmatory factor analysis (see, for example, https://doi.org/10.1207/s15327906mbr3302_1 and https://doi.org/10.1002/hrm.21852), which was not a problem in our case. Moreover, we also wanted to balance the comprehensiveness of the items with feasibility, since particularlyin the context of COVID-19 it is helpful to have scale that is quick to complete so that it can bedeployed more easily and is more likely to be completed by respondents. Indeed, there has been a trend towards favoring short and even single-item scales in part for their added feasibility. Moreover, in cases where inter-item correlations are high (such as for vaccine hesitancy, where the items correlated strongly at 0.93 - see comment 9 below), the added value of an additional item may be negligible. There are also several examples of clinically meaningful valid scales that include only 2 items in a domain, such as the Pittsburgh Sleep Quality Index (PSQI), the brief big-five measure, (10.1016/0165-1781(89)90047-4 , 10.1016/S0092-6566(03)00046-1 , for example, though of course if feasibility were not a concern then more items would generally be preferable – see, for example, https://doi.org/10.1207/s15327906mbr3302_1). Moreover, the trend towards single item scales (despite validity concerns) has been driven largely by feasibility concerns coupled with demonstrations that shorter scales (and even single-item scales) often correlate closely with their more lengthy counterparts. This has been seen in areas as disparate as psoriasis severityassessment (where the lengthy PASI is probably no better than a single-question assessment),depression (10.1192/bjp.174.3.266), and health quality (https://doi.org/10.2466/pr0.1991.69.1.127). Thus the low number of items in some of our domains was driven by both the clinical narrowness of those domains and a desire for feasibility. If the reviewer and editors feel strongly that we need to collapse items in order to create larger numbers of items per (broader) domain, this is of course possible, but given thatthe current domain menu is both statistically sound and clinically meaningful, we would prefer to leave the domains as they are. Moreover, as seen from the domain intercorrelations(see point 7 below), the domains themselves did not overlap in the EFA so it may not make sense to collapse them together. We have also included a discussion regarding this issue under the limitations section of our manuscript in order to clarify the issue further.

Specific:

(1) I'd like to read a bit more about the domains, in particular whether domains emerged thatwere not considered further (and if so, why). Second, the domains are obviously related. For example, stress, fear, and anxiety could easily be merged into a single domain, so I wonder what was the rationale to keep them apart.

Thank you for your important question. The main domain that we considered initially that was ultimately rejected was anxiety; as noted in the manuscript, we initially assumed that anxiety would be an important standalone domain, but ultimately it had significant overlap with stress (and unifactorial loading); we note in the manuscript that "Anxiety as a standalone domain was culled due to unifactorial loading and significant overlap with the more general stress domain." We did consider combining stress and fear, but given that weconsidered them as distinct domains a priori (and this is consistent with prior work, such as https://dx.doi.org/10.1016%2Fj.janxdis.2020.102232) and that they loaded as distinct domains on the EFA as well, we felt that we had sufficient clinical and statistical grounds tokeep them as separate domains. As a general approach, we also felt that for researchers working with this scale it would be better to err on the side of over- rather than under- dividing the domains to make our scale as useful as possible to researchers.

(2) How and how many items were generated for each domain?

A total of 50 items were generated through an iterative process based on the existing literature and pilot study responses; we initially considered 8 domains, with 4-10 items included per domain depending on the clinical breadth of the domain. We have added a sentence to this effect as well: "We included between 4 and 10 items per domain a priori for consideration, depending on the clinical breadth of the domain."

(3) Though I doubt that results will change, the decision on the number of EFA factors to retain should nevertheless be made in line with recent recommendations, 10.1037/met0000200 . I'd also like to see the largest eigenvalues.

Thank you for providing the reference to your recent excellent and comprehensive manuscript. We have now included a reference to your work in our manuscript, and clarify the importance of using multiple approaches in establishing the appropriate number of EFA factors to retain. We have also included the eigenvalues (3.97, 2.49, 1.69, 1.12, 1.40, 1.25, and 1.48) and this has been added to the manuscript.

(4) As the authors mention themselves, it does not appear reasonable to drop items based on low interitem correlations, simply because some domains (say fatalism) may be independent of the other dimensions, so we would not expect any meaningful cross-domain correlation.

Thank you for highlighting this important point. We have added language to the manuscript outlining that such deletions were only performed when it made clinical sense, rather than based on low correlations alone.

(5) Convergent/divergent validity was merely assessed through the relations to a single item each, so these results should not be overstated. I recommend something along the lines of "preliminary validity evidence".

Thank you for raising this important point. We agree absolutely and have further clarified this in our limitations: "Second, further convergent and discriminant validity assessments could be considered using other established scales; since the OPAS-C is unique in its broad multifactorial structure, however, this would require the inclusion of multiple scales and additional questions. Moreover, the convergent and discriminant validity seen using single-question comparators support the validity of the OPAS-C. Still, since these rely on single- question comparators this can be regarded as preliminary validity evidence.

(6) p.5: "items were deleted due to low polychoric and polyserial correlations, unifactorial outcomes, duplication, or poor correlations.". The item-reduction process is unclear. What is meant by "poor correlations" and how does that differ from "low...correlations"? What are "unifactorial outcomes"? And why were item-duplicates considered at all?
I suggest to remove items when they exhibited very strong correlations to another item (duplicates) and afterwards just consider the factor loadings and maybe endorsement rates.

Thank you for raising these points; you are absolutely correct that we were redundant in our discussion of redundancy, and we used the term duplication rather than redundancy in our initial version of the manuscript. Unifactorial outcomes was meant to refer to a situation where a single item alone maps to a factor. We apologize for the lack of clarity in our initial wording and have clarified this in the manuscript.

(7) Please report the factor intercorrelations.

We have added Table 3 to outline the factor intercorrelations. The clinical meaningfulness of the factor structure is further highlighted by this matrix, though it also suggests that the decision to keep the factors separate, rather than collapse them, together, is reasonable.

	I	II	III	IV	V	VI	VII
I	1						
II	0.436	1					
III	0.369	0.474	1				
IV	0.138	0.298	0.461	1			
V	0.309	0.303	0.441	0.399	1		
VI	0.277	0.412	0.590	0.462	0.382	1	
VII	0.315	0.257	0.236	0.420	0.228	0.376	1

(8) Just report the robust SEM results.

In Stata, robust SEM results include only the SRMR (Stata argues that the other goodness of fit indices are not appropriately reported with robust SEM, though other statistical packages such as MPlus do indeed report them with robust SEM); because we felt that some readers may be curious regarding the other goodness of fit indices, we thought it would be useful to include those for reference, though we of course defer to the reviewers and editors on this matter.

(9) I doubt a "total OPAS-C score" makes sense. If there are 7 domains, what should the total score express? Relatedly, alpha for the full scale does not seem to make sense either, so please report reliabilities for each subscale.

Thank you for raising this important issue. The overall score is designed to measure the degree of distress/ undesirable attitudes towards the pandemic. The question of whether the overall score should be used as a composite measure and the relative value of the subscales is an important one, and this is an area of future research as we explore the most clinically meaningful ways in which to use this scale. While we would not claim that any attitude is "wrong" or dysfunctional, it will be interesting to see the ways in which the validity of this scale will be borne out in future studies where we aim to use the OPAS-C as a composite measure. The alpha for the full scale (0.87) appears adequate and echoes the clinical meaningfulness, particularly since this alpha remained unchanged for both the US and UK samples. In terms of subscales, the alpha values were 0.90, 0.38, 0.82, 0.66, 0.77, 0.76, and 0.93; not surprisingly, subscales with smaller numbers of items had lower alphas but this would be expected.

(10) Please be more precise concerning the multigroup model. Was it a strong invariance model or just a configural invariance model? In either case, either provide the full sequence of invariance testing or just drop these analyses.

Thank you for asking about the multigroup model. We used scalar (strong) invariance, as we held the means (intercepts) constant across the two groups—that is, we held the structural coefficients and structural intercepts constant across groups. This approach was clarified in the manuscript; see also, for example, https://dx.doi.org/10.1016%2Fj.dr.2016.06.004.

(11) How it comes that there's a loading > 1?

We used Promax oblique rotation, and therefore the factor loadings represent regression coefficients rather than correlation coefficients and can sometimes be larger than 1. See, for example:
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjZn8L
DuIDuAhXltlkKHYuIAeEQFjAAegQIBBAC&url=http%3A%2F%2Fwww.statmodel.com%2Fdownl
oad%2FJoreskog.pdf&usg=AOvVaw3voEtZiIysbZqkVjGQKW3v
Minor:

- Abstract: "The population" => the sample.

Thank you. This was changed in the abstract.
- Table 1: just report two decimals and add factor labels

Thank you for this suggestion. The table was amended as you suggested.

REVIEWER	Moshagen , Morten
	Ulm University
REVIEW RETURNED	22-Jan-2021

GENERAL COMMENTS	The authors have been response to my comments. I think that the manuscript has substantially improved. I have a couple of further comments:
	Regarding the robust SEM results: Is is incorrect that fitindices other than the rmsea are not properly defined when performing robust tests. If stata does not report these, the authors could simply use any other sem software, such as lavaan in R, which is freely available. Non-robust indices of fit are simply not meaningful in the presence of non-normality, so there is no point in reporting these. Relatedly, the statement "When performed with robust standard errors, the SRMR was 0.056," does not make sense, given that using "robust standard errors" just corrects the standard errors, but not necessarily model fit statistics (though usually both standard errors and test-statistics are corrected, but these are actually different issues). So, please (a) only report robust sem results along with all relevant indices and (b) explicitly state which estimator and correction procudure has been employed.
	The results on the invariance tests are still not informative. In the methods section, the authors write "holding means and covariances equal across groups", which would actually imply equality of covariance matrices and means, however, in the response letter the authors stated that thy "held the structural coefficients and structural intercepts constant", which implies a strong (scalar) invariance model. The latter makes actually more sense, to please be precise. However, it is unwise to test for strong (scalar) invariance, without determining weak (metric) invariance first. So please either compute and report the usual sequence along with difference statistics, i.e., (1) configural, (2) metric, (3) scalar invariance, or simply drop these analyses. See e.g. 10.1080/17405629.2012.686740 for a simple introduction to measurement invariance.
	Regarding a total score: I maintain that a total score is not meaningful given that there are 7 rather weakly related domains. If the authors want to make the point that there is something akin to "degree of distress/ undesirable attitudes towards the pandemic" and that the subscales represent more narrow facets, they need to back up this claim by using an appropriate statistical model, say a higher order factor or a bifactor model.
	With respect to the loading > 1. Yes, standardized regression can be larger than one, however, when factor loadigns are > 1, this implies a negative variance estimate for the respective indicator and thus indicates that the hypothesized model is wrong. Something needs to be done about this.

Reviewer: 2

Comments to the Author:

The authors have been response to my comments. I think that the manuscript has substantially improved. I have a couple of further comments:

We thank the reviewer for his kind compliment, and appreciate that the improvement in the manuscript is largely due to his helpful and constructive comments.

Regarding the robust SEM results: Is is incorrect that fit indices other than the rmsea are not properly defined when performing robust tests. If stata does not report these, the authors could simply use any other sem software, such as lavaan in R, which is freely available. Non-robust indices of fit are simply not meaningful in the presence of non-normality, so there is no point in reporting these. Relatedly, the statement "When performed with robust standard errors, the SRMR was 0.056," does not make sense, given that using "robust standard errors" just corrects the standard errors, but not necessarily model fit statistics (though usually both standard errors and test-statistics are corrected, but these are actually different issues). So, please (a) only report robust sem results along with all relevant indices and (b) explicitly state which estimator and correction procudure has been employed.

We thank the reviewer for his helpful comments, and we concur absolutely with his opinion that there is no point in reporting the non-robust indices of fit simply because of a limitation in Stata. These were deleted, as suggested. We then used Lavaan, as suggested by the reviewer, to run the analyses, and a full array of fit indices (using maximum likelihood estimation with robust standard errors and a Satorra-Bentler scaled test statistic) have now been included in the manuscript, and we have only included robust indices of fit.

The results on the invariance tests are still not informative. In the methods section, the authors write "holding means and covariances equal across groups", which would actually imply equality of covariance matrices and means, however, in the response letter the authors stated that thy "held the structural coefficients and structural intercepts constant", which implies a strong (scalar) invariance model. The latter makes actually more sense, to please be precise. However, it is unwise to test for strong (scalar) invariance, without determining weak (metric) invariance first. So please either compute and report the usual sequence along with difference statistics, i.e., (1) configural, (2) metric, (3) scalar invariance, or simply drop these analyses. See e.g. 10.1080/17405629.2012.686740 for a simple introduction to measurement invariance.

Thank you for your helpful comments and insights. We also thank the reviewer for the excellent reference on measurement invariance. We have chosen to drop these analyses, as suggested by the reviewer as an option, in order to simplify the manuscript for the non-specialist reader.

Regarding a total score: I maintain that a total score is not meaningful given that there are 7 rather weakly related domains. If the authors want to make the point that there is something akin to "degree of distress/ undesirable attitudes towards the pandemic" and that the subscales represent more narrow facets, they need to back up this claim by using an appropriate statistical model, say a higher order factor or a bifactor model.

Thank you for highlighting the limited value of the total score. This is a critical point, and instead of focusing on the total score we have now added an additional table (Table 4) that includes the reliability for each subscale, as the reviewer suggested.

With respect to the loading > 1. Yes, standardized regression can be larger than one, however, when factor loadigns are > 1, this implies a negative variance estimate for the respective indicator and thus indicates that the hypothesized model is wrong. Something needs to be done about this.

Thank you for highlighting this important point. We have deleted the item that included a loading > 1 from the model given concerns over a negative variance estimate.

Thank you again for considering our work and for your very helpful and constructive comments.