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)	Method Effects Associated with Negatively and Positively Worded Items on the 12–Item General Health Questionnaire (GHQ–12): results from a cross-sectional survey with a representative sample of Catalonian workers
AUTHORS	Rodrigo, Maria F.; Molina, J. Gabriel; Losilla, Vives, Jaume; Josep-Maria; Tomás, José

VERSION 1 – REVIEW

REVIEWER	Jakob Bue Bjorner University of Copenhagen
REVIEW RETURNED	20-Jun-2019

GENERAL COMMENTS	The question of method effects continues to be of interest in questionnaire research. The current study examines method effects in the General Health Questionnaire, GHQ-12, using structural equation models (SEM) for categorical data and a random sample of 3,050 employees from Catalonia, Spain. The sample is adequate and the statistical methods are well suited to the problem. The writing is clear. The presentation of descriptive data for the GHQ-12 items, model fit details, and parameter estimates is adequate. I have some comment on terminology, model specification, and the interpretation of results. 1. The SEM models used are described as correlated traits-correlated methods (CTCM) and correlated traits-correlated uniqueness (CTCU) models. I am not familiar with this terminology. I suggest you provide a bit a description on page 5 when the models are first introduced. 2. As you describe the models later, the CTCM models are also known as bifactor models. I suggest that you refer to the large recent literature on these models (e.g. 1). 3. Your final model 6 specifies a general factor and two factors named method factors, for positive wording and negative wording, respectively. I am concerned that you allow these two factors to be correlated. As have been discussed in recent papers (e.g. 2), allowing method factors to be correlated may introduce problems of model identification, in particular when no item loads only on the general factor – as in your model. I suggest that you also evaluate a model where the method factors are specified as uncorrelated and evaluate the change in item loadings compared to the current model 6. 4. Table 2 seems to be missing the data for model 5.

5. While model 6 has excellent fit, some of the parameter estimates cause concern for your interpretation of a general factor and two method factors. Eight out of 12 items have higher loading on the “method” factors than on the general factor. Based on a rough calculation (see 1), the general factor only explains 32% of the total and 52% of the common variance, which does not suggest a strong general factor. For item 3 (“useful part in things”) and 4 (“capable of making decisions”), the loadings on the general factor are 0.09 and 0.22, respectively. Thus, according to your interpretation, the responses to these two items are almost exclusively due to method effects. For a factor reflecting the effect of positive or negative item wording, I would expect the item loadings to be similar across all items in the same factor. This is roughly the case for the positive wording factor, although the loadings are higher than I would like to see in a method factor. However, the loadings on the negative wording factor show large variation with the strongest loadings (0.83 and 0.72) for items 10 (“Losing confidence in yourself”) and 11 (“Thinking of yourself as a worthless person”). A likely (albeit post-hoc) explanation is that a potential negative wording factor is confounded with a confidence/self-image factor. Due to these considerations, I suggest you do the following:
a. Include two- and three-factor multifactor models in your model comparisons. I would use the models suggested by earlier research.
b. Compare and discuss models, not only in terms of fit, but in terms of parameter estimate. You could compare all models or restrict comparisons to the models with best fit, among the currently presented models, e.g. model 1 and model 6.

6. You write on page 11: “Moreover, the statistically significant correlation between the two method factors (r = .20) suggests … that respondents susceptible to negative method effects are also susceptible to positive method effects.” The meaning of this statement is not clear to me. Presumably, all GHQ items are scored so a high score indicates more health problems. Thus, a high score on a negatively worded item indicates a more than usual occurrence of a negative health state while a high score on a positively worded item indicates a less than usual occurrence of a positive health state. Consequently, a high score on the negative method factor indicates a tendency to indicate more severe health states on the negative items for a given general health level (as indicated by the general factor). A high score on the positive method factor indicates a tendency to indicate more severe health states (less positive experience) on the positive items for a given general health level. You state that the negative method factor and the positive method factor are positively correlation. I find that hard to reconcile with the interpretation: “respondents susceptible to negative method effects are also susceptible to positive method effects.” The example also highlights that the interpretation of the method factors is hard to separate from the interpretation of the general factor when the method factors are allowed to correlate. Please revise the discussion.

Minor comments:
I suggest that you provide more descriptive information about the sample, such as the distribution on gender, age, and education level. Data on job type would also be useful if you have it.

	Page 4 line 32. Referencing Figure 1 model 3 is not so helpful, since Model 3 is not a three-factor model. I suggest you include the described two- and three-factor models in figure 1. Same comment for line 37. Page 13. “Thus, it follows that the respondents assessed their psychological health more positively when answering NW items than when answering PW items.”. Well you sort of clarify this in the subsequent comments, but responses to the positive and negative items cannot be compared since they use different response choices. Please revise. 1. Reise, SP. The rediscovery of bifactor measurement models. Multivariate behavioral research, 2012, 47.5: 667-696. 2 Markon, KE. Bifactor and hierarchical models: Specification, inference, and interpretation. Annual review of clinical psychology, 2019, 15: 51-69.

REVIEWER	Jesús M. Alvarado Facultad de Psicología. Universidad Complutense de Madrid Spain
REVIEW RETURNED	24-Jun-2019

GENERAL COMMENTS	482/5000 The authors must identify the model chosen as a bifactor model, and consequently discuss the problems of this type of models (for example, overfitting). The advantage of a bifactor model is to adequately decompose the variance, which allows to better know the reliability of the instrument. Since a problem of invariance by gender has been observed, it should be investigated in greater depth. A multi-group CFA could be used against the more limited MIMIC approach. Is it a partial invariance problem?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Jakob Bue Bjorner

Institution and Country: Optum

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below The question of method effects continues to be of interest in questionnaire research. The current study examines method effects in the General Health Questionnaire, GHQ-12, using structural equation models (SEM) for categorical data and a random sample of 3,050 employees from Catalonia, Spain. The sample is adequate and the statistical methods are well suited to the problem. The writing is clear. The presentation of descriptive data for the GHQ-12 items, model fit details, and parameter estimates is adequate. I have some comment on terminology, model specification, and the interpretation of results.

1. The SEM models used are described as correlated traits-correlated methods (CTCM) and

correlated traits-correlated uniqueness (CTCU) models. I am not familiar with this terminology. I suggest you provide a bit a description on page 5 when the models are first introduced.
ANSWER: As requested by the reviewer the terminology, and implications of the CTCM and CTCU models have been introduced on page 5.

2. As you describe the models later, the CTCM models are also known as bifactor models. I suggest that you refer to the large recent literature on these models (e.g. 1).
ANSWER: CTCM model may or may not include correlated methods, whereas the bifactor model includes a general trait factor and other (uncorrelated) substantive (or method) factors. As such, bifactor models are a subsample of the CTCM. We have explained this in the place the reviewer has requested and we have offered the proper literature.

3. Your final model 6 specifies a general factor and two factors named method factors, for positive wording and negative wording, respectively. I am concerned that you allow these two factors to be correlated. As have been discussed in recent papers (e.g. 2), allowing method factors to be correlated may introduce problems of model identification, in particular when no item loads only on the general factor – as in your model. I suggest that you also evaluate a model where the method factors are specified as uncorrelated and evaluate the change in item loadings compared to the current model 6.
ANSWER: We agree with all the considerations of the reviewer on model 6. We also agree that a model similar to model 6 but with no correlation between the two method factors would allow for further comparisons. Thus, we have added this model into the list of tested models. Indeed, this model has been the best fitting model when parsimony is considered.

4. Table 2 seems to be missing the data for model 5.
ANSWER: Model 5 is not identified due to the large number of correlated uniqueness. Model 5 was only included by completeness and for being systematic. Nevertheless, as we understand that including a model that it is not testable may introduce noise for the reader, we have deleted this model from figure 1. Additionally, we have added in the figure the additional models tested (as requested by the reviewer).

5. While model 6 has excellent fit, some of the parameter estimates cause concern for your interpretation of a general factor and two method factors. Eight out of 12 items have higher loading on the “method” factors than on the general factor. Based on a rough calculation (see 1), the general factor only explains 32% of the total and 52% of the common variance, which does not suggest a strong general factor. For item 3 (“useful part in things”) and 4 (“capable of making decisions”), the loadings on the general factor are 0.09 and 0.22, respectively. Thus, according to your interpretation, the responses to these two items are almost exclusively due to method effects. For a factor reflecting the effect of positive or negative item wording, I would expect the item loadings to be similar across all items in the same factor. This is roughly the case for the positive wording factor, although the loadings are higher than I would like to see in a method factor. However, the loadings on the negative wording factor show large variation with the strongest loadings (0.83 and 0.72) for items 10 (“Losing confidence in yourself”) and 11 (“Thinking of yourself as a worthless person”). A likely (albeit post-hoc) explanation is that a potential negative wording factor is confounded with a confidence/self-image factor. Due to these considerations, I suggest you do the following:
a. Include two- and three-factor multifactor models in your model comparisons. I would use the models suggested by earlier research.
b. Compare and discuss models, not only in terms of fit, but in terms of parameter estimate. You

could compare all models or restrict comparisons to the models with best fit, among the currently presented models, e.g. model 1 and model 6.
ANSWER: All considerations by the reviewer are plausible, interesting and should lead to the proposed changes. According to a literature review on multidimensional models for the GHQ (Tomás et al., 2017), the three factor model proposed by Graetz (1991) has the greatest support. Therefore, and following your advice we have included this model plus the three-factor model with method effects. These models did not improve model fit of the one-factor model with method effects associated to PW and NW worded items, but allowed to clarify that method effects were not simply a product of shared substantive variance. We have discussed best-fitting models not only in terms of global fit, but also regarding parameter estimates.

6. You write on page 11: “Moreover, the statistically significant correlation between the two method factors (r = .20) suggests … that respondents susceptible to negative method effects are also susceptible to positive method effects.” The meaning of this statement is not clear to me. Presumably, all GHQ items are scored so a high score indicates more health problems. Thus, a high score on a negatively worded item indicates a more than usual occurrence of a negative health state while a high score on a positively worded item indicates a less than usual occurrence of a positive health state. Consequently, a high score on the negative method factor indicates a tendency to indicate more severe health states on the negative items for a given general health level (as indicated by the general factor). A high score on the positive method factor indicates a tendency to indicate more severe health states (less positive experience) on the positive items for a given general health level. You state that the negative method factor and the positive method factor are positively correlation. I find that hard to reconcile with the interpretation: “respondents susceptible to negative method effects are also susceptible to positive method effects.” The example also highlights that the interpretation of the method factors is hard to separate from the interpretation of the general factor when the method factors are allowed to correlate. Please revise the discussion.
ANSWER: The reviewer is right that interpretation of this correlation is hard in the context of method effects. Nevertheless, as we have included a bifactor model (with independent method factors) and this model had the best fit, the problem with this interpretation no longer exists. We have changed discussion accordingly.

Minor comments:
I suggest that you provide more descriptive information about the sample, such as the distribution on gender, age, and education level. Data on job type would also be useful if you have it.

ANSWER: We have included the following table in the manuscript.

	Mean (SD)	n (%)	Range
Gender			
Women		1361 (44.6)	
Age	40.46 (11.19)		17-82
Education			
Incomplete primary studies		90 (3.0)	
Primary studies		541 (17.9)	
Secondary studies. 1st stage		637 (21.0)	
Formació professional		763 (25.2)	

High School	598 (19.8
Graduate studies	359 (11.9)
Postgraduate studies	39 (1.3)

Page 4 line 32. Referencing Figure 1 model 3 is not so helpful, since Model 3 is not a three-factor model. I suggest you include the described two- and three-factor models in figure 1. Same comment for line 37.

ANSWER: Now all models are presented both as figures and also explained in page 7. Model numbers, 1, 2, 3, etc. are only used for easiness of notation.

Page 13. “Thus, it follows that the respondents assessed their psychological health more positively when answering NW items than when answering PW items.”. Well you sort of clarify this in the subsequent comments, but responses to the positive and negative items cannot be compared since they use different response choices. Please revise.

ANSWER: This has been revised according to the new models tested and results of the best fitting model.

1. Reise, SP. The rediscovery of bifactor measurement models. Multivariate behavioral research, 2012, 47.5: 667-696.
2 Markon, KE. Bifactor and hierarchical models: Specification, inference, and interpretation. Annual review of clinical psychology, 2019, 15: 51-69.

Reviewer: 2
Reviewer Name: Jesús M. Alvarado

Institution and Country: Facultad de Psicología.
Universidad Complutense de Madrid
Spain

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
482/5000
The authors must identify the model chosen as a bifactor model, and consequently discuss the problems of this type of models (for example, overfitting).
ANSWER: According to requests by reviewer 1, new models have been introduced. Among them a bifactor model with a general health factor plus two method factors associated to NW and PW items that are uncorrelated. The best fitting model in the previous version of the manuscript was a CTCM model with correlated method. Nevertheless, in the new version it is the bifactor model that has better fit. Therefore, the bifactor model has been discussed.

The advantage of a bifactor model is to adequately decompose the variance, which allows to better know the reliability of the instrument.

ANSWER. The reviewer is true. Nevertheless, the aim of the manuscript is to address method effects in the GHQ

Since a problem of invariance by gender has been observed, it should be investigated in greater depth. A multi-group CFA could be used against the more limited MIMIC approach. Is it a partial invariance problem?
ANSWER: We have not studied invariance by sex, age or educational level. What we have done is to stablish a MIMIC model to test for potential effects of these covariates on the method factors, or in other words, if method bias were more likely depending on sex, age and/or educational level. MIMIC models are adequate for solving these questions (see, for example, Thompson & Green, 2015). Measurement invariance by sex, age and/educational level was far beyond the scope of the manuscript.

Thompson, M. S. & Green, S. b. (2015). Evaluating Between-Group Differences in Latent Variable Means. In G. r. Hancock and S. O. Mueller (Eds.), Structural Equation Modeling: A Second Course (2nd edition), pp 163-219. NC, Charlotte: INFORMATION AGE PUBLISHING, INC.

VERSION 2 – REVIEW

REVIEWER	Jakob Bue Bjorner University of Copenhagen
REVIEW RETURNED	05-Aug-2019

GENERAL COMMENTS	Thanks for your responses. The new model 7 is easier for me to interpret than the previous models. Please find some followup on the previous comments below. The numbers refer to my original comments: Re. 1-2. Description of CTCM and CTCU models and discussion of bifactor models. These models are now better described and the inclusion of bifactor models is fine. However, the text on page 6 could be streamlined further to avoid repetitions and have a better flow. Also, there are some wording problems in the text, e.g. line 21: "… can be explained by can be written as…". Please revise to eliminate errors and increase readability. Re 5. Thank you for including two- and three-factor multifactor models in your model comparisons. The comparison clarifies that these models do not provide notably improved fit. However, I still think you need to discuss the interpretation of parameter estimates. While you now interpret a bifactor model, some of the issues mentioned in my previous review are still pertinent. Eight out of 12 items have higher loading on the "method" factors than on the general factor. The general factor still only explains 32% of the total and 52% of the common variance. For item 3 ("useful part in things") and 4 ("capable of making decisions"), the loadings on the general factor are 0.09 and 0.14, respectively. Thus, according to your interpretation of the two local factors as method factors, the responses to these two items are almost exclusively due to method effects. For a factor reflecting the effect of positive or negative item wording, I would expect the item loadings to be similar across all items in the same factor. This is roughly the case for the positive wording factor, although the loadings are higher

	than I would like to see in a method factor. However, the loadings on the negative wording factor show large variation with the strongest loadings (0.70 and 0.72) for items 10 (“Losing confidence in yourself”) and 11 (“Thinking of yourself as a worthless person”). A likely (albeit post-hoc) explanation is that a potential negative wording factor is confounded with a confidence/self-image factor. Due to these considerations, I suggest you provide further discussion of the parameter estimates. If your interpretation of the two local factors as methods factors is correct, is it still reasonable to interpret the overall score, given the magnitude of the methods factor loadings? Re 6. The discussion of the relation between factors and sociodemographic variables have been revised and simplified. However, further clarification would be helpful. For example, the scoring of the sex variable is unclear to me. However, if we assume that men are scored 0 and women scored 1, the results would suggest that women has a worse overall score, but this effect is partly modified on scale score level by a methods effect on the positively worded items, where women tend to indicate better psychological general wellbeing than indicated by their score on the global factors. The results for age seem to suggest that older respondents have worse general psychological well-being and this effect is magnified by a methods effect on the positive wording factor. Please provide clarification in these sections on page 13. Minor comments: There seems to be some grammar problems in the new bullets on study strengths and limitations. Please correct. I suggest writing: ”- Comparison of confirmatory models for positively and/or negatively worded items and the use of two different parameterizations. - Investigation of demographic correlates of wording effects”

REVIEWER	Jesús M. Alvarado Departamento de Psicobiología y Metodología de las Ciencias del Comportamiento. Facultad de Psicología, Universidad Complutense de Madrid. Spain.
REVIEW RETURNED	12-Aug-2019

GENERAL COMMENTS	I believe that the authors have responded correctly to the suggestions of the reviewers.

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1
Reviewer Name: Jakob Bue Bjorner
Institution and Country: Optum, USA
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
Thanks for your responses. The new model 7 is easier for me to interpret than the previous models. Please find some followup on the previous comments below. The numbers refer to my original comments:

Re. 1-2. Description of CTCM and CTCU models and discussion of bifactor models.
These models are now better described and the inclusion of bifactor models is fine. However, the text on page 6 could be streamlined further to avoid repetitions and have a better flow. Also, there are some wording problems in the text, e.g. line 21: "… can be explained by can be written as…". Please revise to eliminate errors and increase
readability.
ANSWER: We have rewritten the text in page 6 regarding the models to eliminate errors and increase readability, as suggested.

Re 5. Thank you for including two- and three-factor multifactor models in your model comparisons. The comparison clarifies that these models do not provide notably improved fit. However, I still think you need to discuss the interpretation of parameter estimates.
While you now interpret a bifactor model, some of the issues mentioned in my previous review are still pertinent. Eight out of 12 items have higher loading on the "method" factors than on the general factor. The general factor still only explains 32% of the total and 52% of the common variance. For item 3 ("useful part in things") and 4 ("capable of making decisions"), the loadings on the general factor are 0.09 and 0.14, respectively.
Thus, according to your interpretation of the two local factors as method factors, the responses to these two items are almost exclusively due to method effects. For a factor reflecting the effect of positive or negative item wording, I would expect the item loadings to be similar across all items in the same factor. This is roughly the case for the positive wording factor, although the loadings are higher than I would like to see in a method factor. However, the loadings on the negative wording factor show large variation with the strongest loadings (0.70 and 0.72) for items 10 ("Losing confidence in yourself") and 11 ("Thinking of yourself as a worthless person"). A likely (albeit post-hoc) explanation is that a potential negative wording factor is confounded with a confidence/self-image
factor. Due to these considerations, I suggest you provide further discussion of the parameter estimates. If your interpretation of the two local factors as methods factors is correct, is it still reasonable to interpret the overall score, given the magnitude of the methods factor loadings?
ANSWER: The reviewer is true that these "almost too strong" method effects are not obvious and may be considered not only method effects but also integrating some shared trait common variance. It is true that NW and PW share method, but probably this shared method variance cannot fully explain the high loadings. We have therefore tempered our discussion and we have added further discussion on these issues in the discussion section, as requested by the reviewer.

Re 6. The discussion of the relation between factors and sociodemographic variables have been revised and simplified. However, further clarification would be helpful. For example, the scoring of the sex variable is unclear to me. However, if we assume that men are scored 0 and women scored 1, the results would suggest that women has a worse overall score, but this effect is partly modified on scale score level by a methods effect on the positively worded items, where women tend to indicate better psychological general wellbeing than indicated by their score on the global factors. The results for age seem to suggest that older respondents have worse general psychological well-being and this effect is magnified by a methods effect on the positive wording factor. Please provide clarification in these sections on page 13.
ANSWER: The scoring of sex is correct (and it has been added in the method section), and therefore the interpretation the reviewer does is also correct. We have added this clarification in the discussion.

Minor comments:
There seems to be some grammar problems in the new bullets on study strengths and limitations. Please correct. I suggest writing:
"- Comparison of confirmatory models for positively and/or negatively worded items and the use of two different parameterizations.
- Investigation of demographic correlates of wording effects"

Reviewer: 2
Reviewer Name: Jesús M. Alvarado
Institution and Country:
Departamento de Psicobiología y Metodología de las Ciencias del Comportamiento.
Facultad de Psicología, Universidad Complutense de Madrid. Spain.

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
I believe that the authors have responded correctly to the suggestions of the reviewers.

VERSION 3 – REVIEW

REVIEWER	Jakob Bue Bjorner Optum Patient Insights
REVIEW RETURNED	20-Sep-2019

GENERAL COMMENTS	I am fine with your responses and revisions. You may consider two minor suggestions for language: Page 7, Line 18, I suggest writing “…are the most popular ones [45], in particular the CFA with correlated traits…” Page 15, line 10, I suggest writing: “… had very low loadings on the…”

VERSION 3 – AUTHOR RESPONSE

We have addressed the editorial request regarding the Strengths and limitations section so that there is one sentence en each point and changed "confusion" to "confounding" variable.
We have also attended the two reviewer 1 suggestions.
We appreciate your suggestions!