|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | Victor Rouco  PhD student  Email: victor.rouco@ugent.be  Faculty of Psychology  H.Dunantlaan 2  B-9000 GENT  Belgium  www.ugent.be |
|  | |  |
|  |  |  |
| date  16/06/2021 | page  1/1 |  | |

Dear Dr Proyer,

We would like to thank you and the reviewers for the constructive comments on our manuscript, entitled, “**The Berlin Multi-Facet Personality Inventory: An IPIP-Based Measure of Big Five Personality Facets”** (Ms number PTAD-21-0007). We hereby enclose our revised manuscript, now consisting of 7,658 words, 29 pages, including 1 Table. We carefully addressed each suggestion that was made and clarify our responses below in a point-by-point fashion.

We are looking forward to any further comments or suggestions, and hope that our manuscript will now meet the standards of Psychological Test Adaptation and Development.

On behalf of all authors,

Sincerely,

Victor Rouco

**Comments by the Editor**

1. **The concern that the study in its present form does not show that the Berlin M-F PI outperforms comparable measures in the field.**

We understand the concern and, although our instrument does offer some other other properties that can provide justification for developing a new instrument (i.e. it is open-source, it has been developed in two different countries) we did not test whether it outperforms other existing measures of Big Five personality. We have now included, nevertheless, a content analysis following the approach of Rosenbusch et al. 2020, which understates the uniqueness of some of our facets while signaling the existing overlaps with other measures. This content analysis is provided now in the supplemental materials uploaded to the OSF page of this project. Furthermore, we included a few sentences in the intro (page 7-8), in the results section (page 14), and in the discussion section (pages 18-19) where we explain the results of this content-wise analysis in the manuscript.

1. **I would also appreciate more information on what suggestions you have regarding the cross loadings and intercorrelations among the facets. I do not think that either of these secondary loadings will be a major problem, but had to be expected.**

As you mention, it is not uncommon to have such cross-loadings in personality data, as it is very difficult to phrase an item that is essentially unidimensional in the sense that it does not tap into other domains. To overcome this issue, we use factor scores through the manuscript, instead of plain sum-scores. The former procedure takes into account the existence of such cross-loadings. We now refer to this in the manuscript in page 20.

1. **A minor additional question I had when inspecting the items was he specificity of some of the facets. For example, “O4; Interest in reading.” Four of the five items contain the word read(ing) and the other one is about enjoying discussing books with others […] In other words, are the items here grouped psychologically meaningful as a facet of openness or is this a reflection of their similarity in the wording?**

Thank you for raising this interesting point. Indeed, the facet that you mention is very specific. However, there is recent evidence which points at this specific facet as a core aspect of Openness (see Trapp & Ziegler, 2019). We therefore have reasons to believe that even though it is a very specific component of the broader domain, interest in reading is a psychologically meaningful aspect of Openness and it does reflect a substantial phenomenon to consider in the assessment of this domain. We now include the above-cited reference in the manuscript and a line explaining the importance of that specific facet, in the discussion, page 19.

With respect to the specificity of other facets, we hope that the content analysis that we have included in this revision could shed light into this matter.

**Comments by Reviewer #1:**   
  
**1. After creating the inventory with all facets they go ahead and test the basic idea to why they have been constructing a new inventory, i.e. that the facets should have higher criterion validity than the factors. With a few exceptions, the facets do not receive this support and in many cases the factor correlates more strongly with the criteria than the facets do.**

We have now included in Table A.3, which presents the results of the nomological network, a comparison of the of the domains versus the facets. There we show that the sum of facets explains more variance than the domain score.

**2. An interesting aspect of this is that the population is, at descriptive level, very different from the American, the latter was representative of the working part of the German population and not undergraduate students.**

We have now included in the limitations section a referral to this (see page 23), which states that even though the descriptives between the two samples differ considerably, the invariance tests show that our scales measure the same constructs similarly across samples.

**3.** **The discussion is a mixture of a discussion of the instrument's value theoretically and practically, and a new review of the psychometric evaluation (which could probably be moved to the result section).**

Our view in this matter is that, given that the PTAD is a journal focused on test development, the discussion of the instrument’s value and its psychometric properties are well addressed in the discussion section of the manuscript, rather than in the results section solely.

**4. The least you can ask for is then that the new test is compared with one of the old ones based on the same item set.**

We have included a new reference in the manuscript where the same taxonomy has been used, albeit with a different set of items, in which the authors test the congruent validity of this taxonomy with relation to other Big Five instruments (page 7). We agree, however, that future research is needed to inspect the congruent and discriminant validity of the BMFPI with other Big Five instruments. We now include this limitation in the limitations section (page 23).

**5. The items of the various facets are almost always keyed in the same direction …**

This is an interesting observation. However, if the negatively keyed facets were only variants of the positively keyed peers, this could be observed either on the residual correlation matrix of the ESEM model or in its Modification Indices. We include both of these in table A.8 and table A.9 in the supplemental materials, only for the sample of study 1 due to simplicity (if the problem that the reviewer points out is true, it would be evident in these tables regardless of the sample used). The largest Modification Indices are not suggested for facets with items keyed in opposite directions, and the correlations of residuals between these pairs are not important. Therefore, there is not any evidence that suggests that negatively keyed facets are simply variants of positively keyed facets.

**6. I cannot see in either the manuscript, the appendix or in the supplement that secondary loadings have been presented.**

We have included the ESEM factor loadings of the two studies in the supplementary materials.

**Comments by Reviewer #2:**   
  
  
**1. Temporal stability reliability is not explored.**

Indeed, we did not test temporal stability in this study. However, a previous study which used the same taxonomy (Ziegler et al., 2019) did explore temporal stability finding satisfactory test-retest reliability. We include this in the limitations section (page 23).

**2. Regarding the nomological network, the authors focus on a limited range of variables related to only some dimensions: Well-being is expected to go mainly with N and E facets, while Academic performance and absentism go mainly with C. However, in order to evaluate the complete nomological network, associations with A and O should also be included as well, for example antisocial behavior or substance use for low A, or political orientation or creativity for O.**

We agree. We did not run a complete test of the nomological network, and this extends to A and O. Future research will hopefully investigate this. We include this issue as a new limitation in page 23.

**3. A central aspect to evaluate the quality of an instrument is the concurrent and discriminant validity. In this sense, and given that the instrument shows relatively new facets, it would be necessary to include another consolidated instrument that would include the dimensions and a wide range of facets of the B5, in my opinion.**

This point raises the same question as point 4 of reviewer 1. We fully agree with this limitation and hope that, by publishing this instrument, others will take up the work and provide more evidence on the congruent and discriminant validity of the BMFPI.

1. **Something that worries me a lot is the replicability of the results. The facets of C that are obtained in the present study do not coincide with the facets extracted in a previous study by the authors using a similar methodology. Indeed, I would have used the CFA not to explore the one-dimensionality of each facet (the internal consistency indices already reflect this), but to explore the replicability of the facets extracted with the EFA.**

We agree that replicability is always an issue that has to be closely examined. We hope that future research could shed light into the degree to which the psychometric properties that we have observed are replicated in other studies. We do have, however, evidence that the instrument’s measurement properties could be robust to future replications, in the sense that we have tested it with two independent samples and that the same taxonomy has been replicated with another instrument.

1. **If I don't misinterpret, the ESEM in Germany shows factor loadings that do not replicate the 5-factor structure found in study 1. In fact, the factor loadings of the German sample seem anomalous (negative loadings, N and O facets presents factor loadings below .30…).**

Even though there are several deviations in the descriptive level, and that some point estimates (i.e. factor loadings) vary with respect to the two samples, we have evidence that a similar structure has been found thanks to the promising results of the measurement invariance tests.

1. **key step in the search for facets is the selection of items, since, as is well known, a factor analysis does not extract anything more than what enters. Therefore, the scale / item selection process should be described in more detail in the manuscript or in the annex.**

The item selection process has been performed empirically by starting from an extensive item pool (the IPIP). We think that we could not add any new substantial details into the manuscript with this regard.

1. **The analyses are sophisticated and adequate, but very confusing. Perhaps they should be explained in much more detail. Is the EFA done with all the items of a dimension in order to extract the facets? And the CFA only for each facet? ESEM is performed at the facet level, not the item level, is that correct? What program has been used to perform the analyses?**

In order to comply with the journal’s space limitations, we were forced to include those details onto the supplemental materials, while describing the statistical analyses in a more high-level fashion in the manuscript.

1. **Despite the large amount of analyses, only one table is presented in the manuscript, and the rest in the annexes. I would also include table 2 and a graph (or table) with the ESEM with the saturations of study 1 and study 2, (saturations of the facets in their dimension, but also in the other dimensions) both in the USA and in Germany in order to examine the structure of the questionnaire in detail.**

We were forced to move all the non-essential tables to the supplemental materials in order to comply with the space limitations provided by the journal. We have included the ESEM factor loadings of study 1 and study 2 in the supplemental materials (Tables A.6 and A.7).

1. **Regarding the invariance, and as the authors emphasize, it is really a problem to have used samples so disparate in age, when this variable can affect the mean scores obtained on each scale, especially C and N (according to the changes in mean scores linked to maturation processes, Roberts et al., 2006). Thus, and according to the other results about invariance, it is inaccurate to conclude that the questionnaire is comparable in both countries.**

We disagree with this point raised by the reviewer. The vast majority of our scales are comparable at the metric level (according with the Measurement Invariance tests), and they are therefore comparable across countries, regardless of the sampling differences with respect to demographic variables.

1. **In addition, the author's state that the partial invariance of the full model cannot be calculated, but is not completely true: it can be calculated using an ESEM within CFA.**

Even though we agree with this concern, we think that such analyses would be beyond the scope of this manuscript. We believe that the manuscript falls within the spirit of this journal, as it is an adaptation of the existing IPIP scales. This effort is therefore a first spark for the proposition of a new instrument, and future research should be conducted in order to test for other complexities and to gather more empirical evidence on its measurement properties.