Reviewer #1: In this manuscript, the authors describe the development and initial validation of the Berlin Multi-Facet Personality Inventory (BMFPI), a hierarchical instrument designed to assess the prominent five-factor model (FFM) of personality. The BMFPI initially was developed in English and then translated into German, such that two different versions of the instrument now are available "to facilitate international usability." (Abstract, p. 1)  
  
The manuscript has some attractive features and the BMFPI shows some potential as a personality measure. Nevertheless, after finishing the paper, I do not believe this manuscript merits publication in a top-tier assessment journal. This evaluation is based on two general considerations. First, the current manuscript fails to articulate a compelling rationale to justify the creation of another hierarchical FFM instrument. As the authors discuss in the Introduction, multiple FFM measures already are available for use. This raises an important question: "Why develop another one?"  
  
I thought the paper got off to a promising start in addressing this question, as the authors stated that their goal was to identify "the maximum possible number of non-redundant narrow constructs below each of the Big Five domains." (Abstract, p. 1) This is a laudatory goal, but the authors failed to convince me that they were successful in achieving it. As discussed on p. 15, the initial item pool simply consisted of 525 items drawn from the International Personality Item Pool (IPIP). The authors say very little about how this pool of 525 items was generated, and they do not provide any evidence establishing that it yielded comprehensive content coverage of each FFM domain. Consequently, it is impossible to know whether the BMFPI actually assesses the maximum number of non-redundant facets within each FFM domain. On the whole, I found this aspect of the scale process to be underdeveloped and not very ambitious.  
  
Second, the scale validation aspect of the paper is extremely limited and falls well below the typical standards expected in a top-tier assessment journal. Study 1 reports criterion validity data based on only seven self-report items (the 5-item Satisfaction with Life Scale, plus single items assessing grade point average (GPA) and academic absences without justification). Study 2 contained no validity evidence whatsoever. The authors do not report any convergent and discriminant validity evidence vis-à-vis other measures, including other FFM instruments (e.g., it would be very useful to know how these facet scales correlate with those contained in other hierarchical FFM instruments). They do not examine convergent and discriminant validity in relation to non self-report data (e.g., informant ratings). They do not assess the stability of BMFPI scores over time. Finally, they do not establish in the incremental validity of this new measure by demonstrating that it provides psychologically meaningful information not already available from existing FFM measures. Simply put, the authors need to collect more validity data.  
  
Given the paucity of validity evidence, it struck me as somewhat premature to translate the BMFPI into German. Put differently, I think it would be a good idea to validate the measure more extensively before creating any derivative versions. In this regard, Smith et al. (Psychological Assessment, 2001) discuss various "sins" that researchers commit in developing short forms. Their specific sin #1 is developing a short form of an insufficiently validated measure. It seems to me that the current authors have committed this same basic sin in developing this German translation before fully validating their measure.  
  
SPECIFIC ISSUES  
  
~~1. On p. 2, the authors speculate on why "many different sets of facets have bee proposed," such that "the number and nature of facets below the Big Five domains is far from being consensual." The biggest single reason is that various test developers arbitrarily decided that each domain should have the same number of facets. Thus, the BFI-2 arbitrarily includes three facets per domain; the HEXACO-PI-R contains four facets per domain; and the NEO and IPIP-NEO-120 each specify six facets per domain.~~ Bullshit  
  
2. I will discuss two issues related to Table 1. First, on p .3, the authors state that certain instruments were omitted from the table "due to limited space." This seems like a very weak rationale to me. Please include all available hierarchical FFM measures in this table.  
  
Second, I found the nomological network portion of the table to be confusing and difficult to follow. For example, it was unclear to me what criteria were used to decide what specific evidence should be listed in the table. Moreover, it was unclear what some of the listed variables actually represent (e.g., what is "Stimulation"?). This aspect of the table either should be clarified or eliminated.  
  
3. On pp. 11-13, the authors review the literature regarding how the Big Five relate to life satisfaction, academic performance, and absenteeism. I found these literature reviews to be unnecessarily complex and cumbersome. The key consideration here is that in each of these areas, there are at least two meta-analyses that examine how the FFM domains relate to these criteria. Heller et al. (Psychological Bulletin, 2004) and Steel et al. (Psychological Bulletin, 2008) both examined how the Big Five relate to life satisfaction (Steel et al. also report some facet-level data). Poropat (Psychological Bulletin, 2009; British Journal of Educational Psychology, 2014) examined how the Big Five relate to academic performance. Salgado (International Journal of Selection and Assessment, 2002) and Cuadrado et al. (Journal of Personality and Social Psychology, 2020) examined how the Big Five relate to counterproductive academic behaviors, including absenteeism. I recommend beginning each section by summarizing these meta-analytic results; these then can be supplemented with facet-level findings from individual studies.  
  
~~4. On p. 12 (and again on p. 40), "Watson and Watson (2002)" should be "Gray and Watson (2002)."~~  
  
5. On p. 15, the authors state that the Study 1 sample "contains part of the sample" reported in MacCann et al. (2009) but "includes more participants." However, the Study 1 sample "consisted of 722 American undergraduate students" (p 14), whereas the participants in MacCann et al. (2009) were students "from a large private high school on the East Coast of the U.S." Please clarify.  
  
~~6. I do not see much value in reporting separate confirmatory factor analyses (CFAs) for each facet. These analyses simply tell you whether a single general factor can account for the correlations among the items (constrained to be a maximum of five) within a facet. They do not provide information regarding whether or not the facet represents a distinct structural unit within a broader hierarchical model.~~  
  
7. Please report the correlations among the domains and facets in the final version of the BMFPI. It is fine with me if these results are reported in a series of supplemental tables.  
  
8. It is noteworthy that many of the Comparative Fit Index (CFI) values reported in the paper (see Table 2 and p. 20) are well below generally accepted guidelines for an adequate fit (on p. 16, the authors themselves cite .90 as the minimum threshold for an adequate fit). Nevertheless, the authors conclude that these model fits were "sufficient" (p. 20) for reasons that are never clearly stated in the paper. This issue merits further discussion in the paper.i  
  
9. Dominance (e.g., "I want to be in charge," "I try to lead others") and Forcefulness (e.g., "I automatically take charge," "I like having authority over others") contain very similar content and seem to assess the same basic trait. Why is the former a facet of Conscientiousness and the latter a facet of Extraversion? Both scales assess content that traditionally is subsumed within Extraversion. This issue needs to be discussed in the paper.  
  
10~~. The Emotional Stability domain is problematic. Six of its seven facets (Equanimity, Confidence, Carefreeness, Drive, Emotional Robustness, Self-Attention) consist entirely of negatively keyed items. Only Mental Balance includes positively keyed items. The domain should be rekeyed and renamed as Neuroticism (or Negative Emotionality or something similar) to reflect the true nature of its assessed content more accurately.~~

It has been manually changed when we named it Mental Balance, negative values are associated with neuroticism.  
  
11. On p. 20, the authors report that they needed to specify correlated residuals between Drive and Persistence, and between Forcefulness and Dominance. These correlated residuals suggest possible discriminant validity problems that need to be further investigated in the paper.  
  
12. On p. 32, the authors acknowledge that they encountered some significant data analytic problems (e.g., model non-convergence). These problems need to be discussed earlier and in greater detail.  
  
  
  
  
  
Reviewer #2: This study sought to develop yet another Big Five measure, the Berlin multi-facetted personality inventory, which is particularly aimed to include the maximum possible number of non-redundant narrow subfeatures of the Big Five domains. As a special quality, the study is carried our in both English and German translations of the measure. The authors conclude that the measure has promising psychometric properties.  
  
The manuscript is well-written, the rationale is sound and innovative, and the sampling and analytical approach generally seems sufficient for the purpose of this study. The overall objective of the study seems unique, innovative, and timely.  
  
The introduction is rather lengthy and ends on page 13, where the aim of the study is presented. However, this might be appropriate for a development study of this kind.  
  
~~There are very few typos - such as on page 13 "predictor absenteeism", where "of" is missing.~~  
  
~~As far as I can see, the authors consistently refer to "The Berlin multi-facetted personality inventory" without using an abbreviation. The authors might consider an abbreviation or acronym for the measure, which can be employed when referring to it.~~

Reviewer #3: PAS-2020-3880  
The Berlin Multi-Facet Personality Inventory: A Comprehensive Measure of Big Five Personality Facets  
  
This manuscript reports findings from two student-sample studies on the development of a new Big 5 measure, with particular attention to facets. A variety of factor and correlational analyses were conducted. It is concluded that the new measure has good structural and external validity, and that the English and German versions of the measure display sufficient invariance.  
  
My impressions of the manuscript are positive, although not highly enthusiastic. On the positive side, I consider the issue to be important. Facets have been a weakness (in my view) in previous measures and in the literature. The new and expanded number of facets are welcome contributions. The analyses were well-done and the findings are straightforward. On the negative side, the external validity evidence (as in correlations with other measures) is rather modest. The problem is not fatal, just underwhelming.  
  
Table 1 list the primary facets from previous measures. The authors correctly note that the "labels vary, they have a substantial degree of similarity in terms of content", and that "Some of these constructs are explicitly covered (e.g. Anxiety), while others are implicitly encompassed".  
  
The issue in the literature is that it is difficult for readers to determine what is the same and what is different across measures, given the varying labels and the only sometimes-varying item contents. The new measure is based on the authors' new, particular selection of items from the IPIP pool. Readers are unable to tell how similar or different the items are from existing measures. But a new array of facets is proposed and they seem as worthy as those in other measures.  
p. 15: "IPIP items were rated regarding their prototypicality for a Big Five domain. Based on these ratings, items for the current study were selected. More details can be found in MacCann et al. (2009)."  
  
Readers need more information about this important part of the item selection procedure. How were inclusion/exclusion the decisions made?  
  
~~p. 17: "ESEM was the preferred procedure as it allows to relax the too strict independent clusters model in which CFA is usually performed".~~  
  
~~Readers should be provided with a more precise description of the benefits of ESEM. The language mistakenly leads readers to believe that it pertains to the degree of correlation between factors, instead of its allowance for cross-loadings.~~  
  
My primary concern is with the many general and overly strong claims that are made about the supposed nomological network and correlations with other measures. These statements or insinuations occur throughout the manuscript, and they are based on correlations with just three other measures: satisfaction with life, grade point average, and class attendance. Consider these relevant statements:  
  
~~p. 17: "Nomological network. In order to examine evidence of construct validity of our proposed facet model, a nomological network linking our constructs with external outcomes was build."  
  
p. 29: "Association with external constructs … Overall, the scores derived here have shown similar test-criterion-correlations as scores from other personality instruments."  
  
p. 31: "All in all, our nomological network sustains a reasonable degree of construct validity for the Berlin Multi-facetted personality inventory."~~  
  
The three external variables do not provide extensive or sufficient external validity correlation tests for the Big 5 or even for most of the facets. The evidence that is presented is quite meager. Agreeableness does not appear in any of the predictions of validity analyses, yet there are 10 new facets of Agreeableness. I think the language needs to be toned down dramatically and that readers need to be made explicitly aware that the presented validity evidence is limited and insufficient.  
  
~~p. 27: "Construct validity was assessed by fitting a CFA to each of the facets."~~  
  
~~The language is imprecise and misleading. At this point, the authors are summarizing their findings on structural validity, which is just one aspect of construct validity.~~  
  
p. 31: "Furthermore, our results suggest that switching the focus from a domain perspective towards a facet perspective may increase the strength of associations between personality scores and external criteria."  
  
I agree. This is a positive feature of the research.