Reviewer's report

Title: Reliability and Validity of the Structured Interview of Personality Organization (STIPO)

Version: 1 Date: 25 March 2013

Reviewer: Kenneth Critchfield

Reviewer's report:

Title: Reliability and Validity of the Structured Interview of Personality Organization (STIPO)

Reviewer's report

- Major Compulsory Revisions

Title: Reliability and Validity of the Structured Interview of Personality Organization (STIPO)

Reviewer's report

This manuscript reports on basic psychometric properties of a German translation of the STIPO interview, based on Kernberg’s theory of personality organization as it relates to other clinical constructs including symptoms, functioning, and personality disorder. The basic objective and sample seem appropriate. However, there are major and pervasive problems with the manuscript that center on a lack of clarity about the underlying theory and how the author’s various methodological choices operationalize and reflect that theory. Clarification of these various issues and decisions (I try to identify many of them below in order of progression through the manuscript) may improve the piece considerably and could merit publication. However, there are also some potential problems that could still remain about whether the overall pattern of findings shows specificity in support of the STIPO’s underlying theory, or whether there may be some general factor, not specifically related to personality organization, that drives the significant findings. The authors will need to present additional data for this to be clear. As noted below, reliability findings also appear to be based on very few observations of cases sampled according to a procedure that is not described, leaving some aspects of this key psychometric property in doubt.

The following consists of notes and comments in order of the manuscript by sections, in hopes that it will be of use in producing a revision. Each of the comments is potentially a major issue, and so all are included in the same reviewer section under “major issues”

Regarding the abstract:
1. The abstract makes no mention of the previous STIPO and IPO work; no references to Kernberg or articulation of the underlying theory. Relevance of the DSM-5 evaluation of personality is unclear. The title and abstract should reflect the fact that this work substantially replicates previous psychometric work with STIPO.

Introduction:

2. The introductory sentence is hard to understand in context. Personality assessment and psychiatric diagnosis certainly have a long shared history prior to DSM-5. Also, the approach put forward by the DSM-5 committee comes from a different set of theories and considerations than those that generated the STIPO. The authors are encouraged to frame commentary about LPFS and DSM-5 closer to recent events. The website provided as a reference is no longer active or publicly available. The DSM-5 committee has chosen to continue PD diagnosis in line with DSM-IV and leave the new approach for further study and development. Also, the Method used in this study is grounded in DSM-IV definitions and does not tackle the DSM-5 issues.

3. The authors present Kernberg's theory with a fair amount of unexplained psychoanalytic terms, along with claims that imply that Kernberg was the first clinician to try and assess personality from a psychoanalytic perspective, which does not seem entirely fair to the history of theory-development and clinical practice in this area. This rhetorical approach will be off-putting to many readers. It would be better for the authors to clearly explain the underlying theory being operationalized, along with details of previous work to develop and use STIPO.

4. It should be noted that “Clarkin and colleagues” work to develop STIPO included Kernberg himself. Other attempts to operationalize personality organisation mentioned by the authors, (e.g., OPD), were not as exclusively interested in operationalizing Kernberg's theory, as the authors imply.

5. The authors provide no reference for the “recently presented” 87-item version of the STIPO, and go on to use a translated 100-item version. Unelaborated mention of a shorter form is confusing, especially since no critique has been provided of the original version (but implies there was some reason to change it).

6. By the end of the Introduction it is not clear what the theory underlying STIPO consists of or what specific predictions flow from that theory. Strengths and weaknesses of the original STIPO and related research with it have not been fully presented (e.g., no mention of the IPO; sweeping generalizations about STIPO's performance in other work without critique). The authors provide no rationale as to why this study is necessary now, how the translation process into German has come about, and so on.

7. As someone who values clear tests of theory, and sees the STIPO as having potential to test Kernberg's particular propositions about it, I find that the Introduction does not make any particular case for how and why this measure is potentially valuable for research and clinical practice – and so it fails to motivate
the need for replication of basic reliability and psychometric findings.

Method

8. The OPD was cited in the Introduction in a manner that implied it was an alternate way of assessing Kernberg's ideas about personality organization – yet, availability of OPD data are withheld from this analysis. This is puzzling since the authors seem to suggest that it would provide good convergent validity evidence, or perhaps represent a competitor model to the STIPO that could be tested for relative ability to predict diagnostic status, etc. Unclear why the OPD was withheld from inspection.

9. Basic rationale for STIPO scoring is underexplained. The authors present a 6-point scale ranging from normality to borderline-ness, apparently as judged by the interviewer (not as a direct combination of item-level scores to the 100 items). It is unclear what rationale led to this decision. Also, what objective clinical meaning does “borderline 1” and “borderline 2” have? Why is psychotic personality organization not included in the scoring labels? What argument is provided that this is not simply a global measure of severity, but actually represents a measure of personality organization as Kernberg envisions it?

10. It is necessary to place some information under Method for how the SCID I and II data are to be used, what predictions are made for correspondence between methods, and so on. Again, articulation of underlying theory and specific expectation for results are lacking/absent.

11. The BPI is described very briefly, including the detail that it is another attempt to operationalize Kernberg's theory. Why and how exactly do they differ? Are these competing operationalizations? How do the authors see their relationship and expect them to perform relative to one another?

Unclear why both SCID-II and the ADP-IV were used together, especially since the authors note the ADP-IV has weaker psychometric properties and is only recommended as a screening device? The results suggest that conclusions from the two instruments diverge, so this is not a minor point.

12. It would help considerably if the authors provide more detail about previous psychometric work with the english-language STIPO, including brief evaluation of it's advantages, disadvantages, and how the current study advances previous efforts.

13. A plan is laid out for convergent validity, but only a weak plan for discriminant validity. The STIPO should have strong correlations with all measures of psychopathology, given its own highly-intercorrelated scale structure (based on other work, it should not be avoided here) and also across instruments. The choice of discriminant validity involving stronger connection with cluster B disorders involves some assumptions about personality theory, as articulated by Kernberg (esp., the idea that a dimension of severity underlies the disorders – not necessarily a feature of DSM diagnosis). This theory and its assumptions should be fleshed out for readers. Regarding discriminant validity per se, are
there any scales expected NOT to correlate significantly in some way? Do the authors predict and test for the greatest correlation of parallel constructs across scales, or simply significant association? Data are not presented under Results to evaluate the possibility of an underlying collinearity or single factor being tested.

14. Need rationale for sampling from two settings, inpatient and outpatient. The outpatient sample consists of referrals to a particular practitioners office, so it is important to describe any possible selection biases that may come about through this narrow entryway into the study.

15. How was training in use of the STIPO conducted? Did the 6 videotapes come from the study sample, or elsewhere? Why was reliability not monitored throughout the study, but only during the training phase? Did interviewers have access to the paper-based data? Did they also conduct the other interviews?

16. Alpha levels look acceptable. Again, what level of intercorrelation existed among scales? Since a total PO score is also used, what are its properties?

17. Do the personality disorder differences on STIPO still hold after controlling for overall symptom severity? It is not clear from these data whether the STIPO is measuring personality organization rather than global symptoms or functioning.

Discussion:

18. The authors seem to over-emphasize reliability findings when compared to other studies that were better designed to test inter-rater reliability properties. The current study was based on only 6 cases and 2 raters watching video. It remains unclear how these 6 cases were selected and whether they were part of the study cases or not. Reliability is a function of the problems/cases rated, training, sampling strategy, and is not just the property of an instrument alone, a fact that further makes claims of 'greater reliability' difficult to make without some description of the different goals, settings, methods, etc.

19. An advantage to the study, underemphasized by the authors, is that they report findings for all 7 scales, not just the subset reported in previous work with STIPO and IPO

20. It is unclear why did the authors not attempt to use the STIPO scales to discriminate among specific types of disorder, but stuck to the broad clusters and a unidimensional notion of severity. The STIPO is assessed here in a quite undifferentiated way along a scale of severity that, as presented, doesn't seem to contribute much that isn't already assessed by other instruments. More clarity is needed about underlying theory, assumptions, and rationale for the specific hypotheses pursued and methods used. The current manuscript is vague on these topics, which weakens the meaning and impact of the various findings.

21. The authors should mention the duration of the interview under method, not just under the discussion.
22. The authors seem to want to talk about the Level of Personality Functioning Scales, but do not describe them in clear terms and have not included them in their study. This emphasis thus seems inappropriate (or is underexplained as to its relevance?).

23. The authors mention a short form, again without citation – but at least make clear in the discussion that the problem has been the length of the instrument. Getting back to the problem of unstated theory, and clinical utility: what are the clinical uses of STIPO and/or Kernberg's structural interview? Does it aid in treatment planning, and if so how? Does it aid in diagnosis, and if so how? So far, it seems only to index the degree of severity or impairment in overall functioning – and takes a lot of effort to achieve - the case for a specific link to personality is less clear. Perhaps the links to theory can help make the case about STIPO's clinical (diagnostic or treatment-related) utility more clear? Without these clarifying links, it is difficult to see the value of these findings.

24. Claims that STIPO has better psychometric properties of inter-rater reliability than OPD (and perhaps SPC) are inappropriate. The authors claim that, for example, OPD diagnosis is only reliable when experts review videotapes without clinical time pressures. However, this is precisely the same paradigm described by the current authors: no attempts were described by the current authors to assess reliability under any other conditions than experts reviewing videotaped interviews. There are other reasons why the comparison of reliability conditions is not apt, including sampling procedures, different scope of the instruments, etc.

25. Another example about why the need to explain underlying theory is important: Only those who already know the STIPO/IPO/Kernberg systems well will understand why “Moral Values” is predicted to have a positive correlation with “Antisocial PD.” Similarly, “Reality Testing” (and other scale labels) actually refer to the reverse construct: “Poor Reality Testing.” Many of the terms and distinctions contained in Table 1 are never explained sufficiently for readers to understand and evaluate the results for themselves.

26. Regarding the Discussion section, there appears to be no detailed review of whether the specific a priori hypotheses were met, or to what degree. There is no discussion of limitations of the study and a tendency to oversell claims about “reliable and valid”

27. Table 2: Why are there no % values for Axis I and II diagnoses? Also, given the nature of the analyses, why aren't N's and %'s given for total PD, as well as Clusters A, B, C? No sense of comorbidity versus “sole” diagnosis, etc.

28. Table 2: the sample characteristics show notable absence of psychotic disorders, schizophrenia, etc. Were there other exclusionary criteria the authors didn't mention? What accounts for the sample compositions in terms of diagnosis? Recruitment only from certain inpatient units? Is this a weakness of the study sample, or does it reflect some choice made by the authors? If a purposeful choice, what might the rationale be?
29. Table 2: It is never clarified how the STIPO “levels” relate to the various subscales. Language of “Neurotic” and “Borderline” with 1s and 2s is left unexplained and not clearly tied to a theory or definition regarding what is being measured.

30. Table 3: The authors seem to avoid typical presentation of an intercorrelation matrix. I suspect this is because there are high intercorrelations among all scales and measures. While it is comforting to see that predicted correlations “panned out” - it would be more illuminating and scientifically defensible to provide the relevant total matrices. The current presentation does not allow readers, for example, to determine whether STIPO “Identity” and BPI “Identity Diffusion” have the highest correlation, or whether the same comparison with other STIPO scales is still higher?

31. Table 3: There appear to be no non-significant correlations in Table 3. Were there any hypothesized scale relations that did not work out? For example, SCID-II dimensional scores for Antisocial PD are not presented in terms of their correlation with STIPO: Moral Values. Instead, only the relevant ADP scale is presented. Why is this the case? Similarly, STIPO scales have subdomains with face-valid connections to other scales, but are not clearly highlighted. For example, the STIPO Aggression scale has “self” and “other” components, as do the comparison instruments. Did the subscale relationships hold up, or no? It is important to know the details to evaluate whether theory is successfully operationalized.

32. Table 4: Again, we are never provided with information about the overlap between STIPO scales. These t-tests in Table 4 are unlikely to be independent. How much of what is driving the effect is shared variance among those scales? How much is simply an index of symptom severity, rather than personality organization per se? What would happen if any attempts were made at statistical control for, perhaps, DSM's GAF (Axis V) rating? (or perhaps demonstrate that another measure of general symptom severity does NOT differ by PD status?). Even if taken as a measure of personality organization, why is a singular dimension deemed preferable to something like the DSM/SCID-II approach? Regression analyses regarding unique versus joint prediction of external variables? Selective presentation of bivariate correlations and multiple t-tests could be consistent with construct validity, but it could also be driven by something far more generic than what the STIPO intends to measure.

33. Table 5: Why were Cluster A patients not included in these analyses? It seems likely they would be informative about the validity of the “reality testing” scale, and perhaps also allow tests of Kernberg's theory of severity towards the Psychotic end of the spectrum. It was mentioned that few “sole” Cluster A patients were available, but reality testing should still be able to be tested among comorbid cases that have these diagnoses.

34. Table 5: Does not clarify in table (as mentioned in text) that all comorbid cases were excluded. Did the authors consider alternatives like retaining all PD
cases and categorizing them by most theoretically-severe disorder? Rationale need to be provided in either case, as both choices have pros and cons relative to underlying theory of the measure. An approach that includes comorbid cases would, as noted above, allow retention of Cluster A for analysis. Also, it would be consistent with the idea that comorbidity is itself a marker of personality disorder severity (uncertain if that particular view is present in Kernberg’s theory, but it has been articulated by other theorists, e.g., Tyrer).

35. Table 5: What is to be made of the lack of differentiation around Coping/Rigidity, and Reality Testing? This is not explicitly mentioned by the authors. Couldn’t it be predicted rather easily that obsessive-compulsive disorders will associate with rigidity, while any presence of Cluster A would correlation with the (poor) reality testing scale? Overall it seems that very low hurdles were presented for these construct-based tests of STIPO.

36. Another use of language that needs to be clarified for readers is Kernberg’s use of “borderline” to mean something different than DSM use of “borderline.” Without it, readers will really be puzzled by lack of any attempt to directly compare DSM-defined BPD with any STIPO scale, instead inspecting only broadly inclusive categories like PD vs. noPD or general Clusters of PD.

- Minor Essential Revisions
- Discretionary Revisions

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.