Reviewer: 1
Comments to the Author
I enjoyed reading the manuscript and applaud the authors for using MTMM techniques in the context of interpersonal communication research. However, I have some reservations about both the assumptions underlying the study and the empirical analysis. I will focus on these issues as I'm not familiar enough with the interpersonal research area to judge the relevance of the topic or the quality of the literature review.

1. My fundamental concern with the study is the confusion of trait and state measures in a single MTMM design. Clearly, both the AELS and NIS measure a generalization of past behavior that is essentially trait-like (i.e. people are good or bad listeners). Moreover, at least on dimension of AELS (sensing) is mostly cognitive, so other-reports have to rely on some social cognition heuristic rather than observed behavior. In this respect, one has to be careful to treat self and other-reports as measures of the same construct.

Partner and observer reports, on the other hand, clearly measure situational constructs, i.e. states rather than traits. While one could expect that generalized listening behavior predicts situation-specific behavior, I do not believe that you could call this a test of mono-trait/hetero-method relationships. I suggest the authors consider treating self/other and partner/observer reports of supportive listening as two distinct constructs. This is also adequate given the study design and the data collection process.

My hypothesis is that the correlations between the two state or trait measures should be higher than state-trait correlations. A cursory glance at table 3 seems to confirm this at least to some extend.

2. I am not convinced that the NIS is (a) unidimensional or (b) valid. Parceling by topic rather than randomly is essentially an acknowledgement that there are, in fact, multiple (7?) subdimensions. And judging from table 4, the authors are even able to properly sort all items and give names for the subscales. As mentioned in the limitations, parceling is a horrible practice when validating scales because information about bad items or lack of convergent/discriminant validity might be hidden by the parcels. Moreover, I cannot see a justification for correlating error terms of parcels except to increase model fit.

Since the authors have already shown they can deal with second order constructs in SEM, why not treat the NIS the same way? My expectation is that the model will still not yield acceptable fit, but then one has to simply concede that the scale is not very good. Moreover, you could actually diagnose items with low or double loadings instead of hiding them in parcels.

3. Still concerning the NIS: Given the items in table 4, I am not sure that a reflective measurement model is adaquate for the scale. The parceling (and also the second order) model assume that all reported behaviors reflect a common immediacy trait or state. I think an argument could be made that NIS (and maybe even situational AELS) is a formative construct, i.e. a composite of different behaviors. This would also explain why NIS is not unidimensional, since

the subscales in formative models do not have to be correlated at all. (see Bagozzi 2011, http://dl.acm.org/citation.cfm?id=2017509)

4. Since the partner was only given a 6-item subset of the NIS, why not use this for self and other as well? As an alternative, and in order to better understand the data, I'd recommend the authors investigate the subscales rather than the full scales in terms of MTMM. This would answer the questions which dimensions of verbal and nonverbal behavior are better suited for hetero-method assessment. Of course, the presentation and interpretation of these MTMM matrices would be somewhat more complicated, but maybe worth the effort.

5. I find the discussion a little verbose and not necessarily to the point. For example, the question of internal consistency is of little importance to the study, and Cronbach's alpha not only assumes tau-equivalence (which was not tested), but tells us little about the reliability or dimensionality of the scales.

6. I'd remove tables 1 and 2 from the paper since they do not really matter for the purpose of the article. Instead, I'd welcome a summary table or path diagram for both scales (so that we can better understand the subdimensions and/or the items used) and maybe another table summarizing the model fit for all 6 CFA.

7. How were MTMM models and the corrections for measurement error (or rather the uncorrected correlations) computed? The method section lacks information about the analytical MTMM strategy and a reference or two. I'm not a real expert here, but there is a large body of literature on using SEM with MTMM data, and there seem to be many choices to make and defend (Eid et al. 2008, DOI 10.1037/a0013219). You cannot expect the readers of HCR to know this approach, so a little more guidance is necessary.

Smaller issues:

1. Mean substitution is not an adaquate strategy for dealing with missing data in SEM since this reduces the variance of the items which is unfortunate for variance/covariance based techniques such as SEM. This does not seem to be a problem for the present study because of very few missings...

2. Are NIS and NIVS the same scales? Then please correct the typo.

3. Please provide a rationale or reference why monotrait-heteromethod correlations should be > .5 or why you expect positive correlations for heterotrait-heteromethod (p.7). On the same page, I think the accepted term for large heterotrait-monomethod correlations is common method bias rather than measurement bias.

Reviewer: 2
Comments to the Author
HCR-12-115
Supportive Listening

Both in the abstract and in the opening pages, I am a little concerned about the (details of) claims that good listening is linked to support. Listening is essential, but the support often comes from what follows from the listening. Someone who is a good listener but blunders in generating support messages will still be ineffective. "Supportive listening" is two behaviors in sequence, in other words, and they need separate operationalizations. "Active listening" seems mainly to be verbal and nonverbal message production (p. 4), rather than reception as is ordinarily understood. You seem to be slipping conceptually between listening that is affirming because the listener seems attentive, and listening that is supportive because it results in support messages. You also seem to be moving without comment between the idea of listening as having perceptively discovered information, and listening as a performance of seeming attentive. Please tidy all this up.

Given the general lit review and the common findings that supportive communication improves health, quality of life, etc., surely the two studies cited on the bottom of p. 5 are outliers. You seem to be trying to make some sort of methodological point here, but without distinguishing the two studies' methods from the rest of the literature, it appears that you are just cherry-picking. It also isn't entirely clear what point you're making – is it just that self-reports don't always match other data sources? That's hardly news, and it isn't even news that self-reports might tend to be self-flattering.

The claim on the top of p. 6, line 4 (that the measurement methods will give different results and that the *differences* will predict well-being) is very strong: it implies that you believe the different measurement methods won't correlate. If they correlated, the differences between their means wouldn't predict anything – they would be just like intercepts in a multiple regression, or a simply way of rescaling self-reports to match observer reports. I don't see justification for this very strong claim.

Surely you need to cite support for the final sentence in the first paragraph to begin on p. 6. On pp. 6-7 you make fairly specific predictions about the size of various correlations without telling the reader what the variables are all going to be, exactly what the measurement techniques will be, and without providing a literature review regarding correlation magnitude. Just predicting that you will get what Campbell and Fiske wanted is not sufficient to justify hypotheses. On p. 15 you say without benefit of citations that AEL and NVI have been found to be orthogonal; this is unsatisfactory.

Too much information about Method is in the footnotes. Please bring it up into text.

On p. 9, you report the fit of a second-order model of AELS without having told us what the first order factors were, what the fit of a first order model was, or why you investigated a second order model at all.

Parceling is controversial in the SEM community (as you eventually point out, some call it 'cheating'), and even when accepted there are various suggestions as to how to form the parcels – randomly, equalizing loadings on the first principal component, etc. It appears that for the NIS you have not so much parceled as you have imposed 7 first order measures on the instrument. You could not have represented all 7 parcels with the 6 items in the NIS-PR. I have real doubts

about this instrument in all its forms, but particularly the PR version. In any event, I think that you have not exactly "confirmed" (p. 16) the NIS measurement model; at best, you have figured out one possibly defensible way to fiddle it.

What statistics were used to assess intercoder reliability? Did you somehow convert something like Cohen's kappa to Cronbach's alpha? Only having 10 data points to assess reliability (active listening) seems insufficient to me. Are the first and third portions of the "behavioral coding" section of Method different, or are they both about one set of codes? You only seem to have two coding projects but seem to have described three. Why did you use Cronbach's alpha after doing CFAs? Why not use one of the CFA reliability measures like AVE, Hancock's H, or another?

On p. 15 you say there are 11 heterotrait-heteromethod correlations, but there are 12. The color coding in the table is not working very well for me, by the way; make sure there is enough textual explanation to make using the colors needless.

Your discussion of the various near-zero correlations offers a variety of possible explanations, but you do not settle on any particular explanation for what you found. This is unsatisfying. It also suggests that you did not design the study to have much possibility of accounting for inconvenient outcomes.

This study offered no particular information about supportive communication, making a good part of the lit review irrelevant and suggesting that the authors have not properly framed what the study was about.

I think the material in Table 3 should probably be published in a journal somewhere. However, I think that there is insufficient theoretical development here to justify publication in a top-tier journal. I suggest that prior to resubmission, the authors reformulate the paper's topic, and supply a literature review that is a lot clearer about the AEL, the NVI, and various substantive results categorized by methodology.