****

**Dr. F. Blake Morton**

Visiting Assistant Professor

Department of Psychology

Franklin & Marshall College, USA

Email: [Morton.blake@gmail.com](mailto:Morton.blake@gmail.com)

Phone: (+1) 252-671-9886

Dear Dr. Bakker:

We would like to thank you for the offer to revise and resubmit our manuscript BEAS-D-15-00445. We would like to thank you, the Associate Editor, and the two anonymous reviewers for providing many helpful commentaries to improve this work. We very much appreciate that you, the Associate Editor, and the reviewers recognize the merits of this paper making a substantial contribution to the literature. As requested, we have revised our manuscript to address all of the comments made. Please find our responses to each commentary below, in boldfaced blue font. We look forward to hearing from you regarding any further queries regarding our manuscript.

Thank you for further considering this manuscript for publication within your journal. We look forward to hearing your decision.

Yours sincerely,

Blake Morton

COMMENTS FOR THE AUTHOR:

Reviewer #1:

I must say that I found the title of the paper and the abstract very interesting. My expectations further increased as I read through the introduction: the brief review of the different methods of the assessment of the number of factors is really good.

**Thank you very much for recognizing these merits of our paper.**

However, the methods and especially the results sections were like an ice cold shadow. They were weak and in a strong contrast with the introduction. The biggest problem with the paper is in my opinion that the results do not provide novel information to the existing literature, lack generality (heavy focused and potentially interesting only to those who study primate social relationships) and the authors' conclusions do not fully agree with their results.

**Thank you for your feedback. We have addressed these and other concerns below.**

1. Major issues

There is a large literature on the assessment of the number of factors. The Kaiser's root-one test has been discouraged many years ago in the psychological and psychometric literature. A short note, bringing attention of animal behaviour researchers to this important issue would be potentially justified. However, this has already been done by Budaev (2010, Using Principal Components and Factor Analysis in Animal Behaviour Research: Caveats and Guidelines, Ethology, 10.1111/j.1439-0310.2010.01758.x).

Indeed, Kaiser’s has a long history of being discouraged, but it is still being used in studies of social relationships. Therefore, one value to our commentary paper is that it brings these issues (again) to people’s attention and/or remind them of the limitation within the context of measuring social relationship structure. While indeed Budaev (2010) examines this issue within a broader context of behavioural studies, Budaev does not examine, report, or discuss the results of PCA structures derived using multiple extraction methods using real data, as our paper does. More importantly, Budaev (2010) does not mention social relationship quality (which is the point of our small commentary paper). Although our paper discusses similar issues as in the Budaev paper (as well as the much broader statistical and psychological literature), our paper is a necessary addition to the literature because it brings such issues into the “spotlight” for studies of social relationship studies. If researchers were aware of the extensive literature on extraction methods, these issues would not be repeated in relationship studies. In other words, our paper will be noticeable to researchers searching specifically for papers that use PCA to measure social relationship quality, which will direct them to the much broader literature on the same issue.

The present results do not provide any significantly new addition, except a much deeper review in the introduction. It would be an excellent study if the authors (a) perform simulations, particularly focusing on small sample sizes typically found in animal behaviour research, (b) supplement such simulation with their real "reference" data and (c) provide a discussion of a wide applicability and interest. However, the data and especially the discussion lack general appeal. No sufficiently general comparisons of the different methods is presented: the authors just applied a few of them to a single (random and perhaps non-optimal) data set.

Extending our reply to the reviewer’s last commentary (see above), we disagree that our paper lacks general applicability and interest. We appreciate the reviewer’s concern, but would like to point out that the purpose of our paper is to provide a “commentary” on previous work (not a literature review or a research article). We use our own data as a case study to aid our discussion – i.e. so readers have something to visualize. The use of PCA to study social relationship quality is relatively new to the literature, and despite there being a very extensive literature on extraction method guidelines, most researchers using PCA to study relationship structure are still making the same mistakes – likely due to the fact that SPSS packages still have scree plots and Kaiser’s criterion set as the default (and only) extraction methods. As we state throughout our paper, we are using data on capuchin monkeys as a case study to highlight our more general commentary on extraction methods used in previous studies of social relationship quality in animals. Again, this approach to studying animal social relationships is still relatively new, and thus far has been applied mainly to primates (and one study of corvids). As we state in our paper, these kinds of analyses are equally useful to studies of other taxonomic groups, but to our knowledge such papers have not yet been published. This makes our paper even more valuable to the literature, because our commentary paper explicitly brings these issues to future researchers’ attention. Indeed the bonobo paper by XXX highlights within their discussion the general value of parallel analysis, our commentary extends this notion by comparing many other robust forms of extraction and – as a commentary paper -- brings these issues directly into the “spotlight” (as opposed to a few commentaries within one’s discussion, which may be ignored or missed by readers).

Furthermore, It is not clearly described how the data were analysed. It seems that the dyads rather than individuals were the units of analysis. As such, data from very few individuals have been pseudo-replicated in the same analysis. How did the authors take care of non-independence between their analysis units? A more complex model should be applied for correct analysis of such data even if such an analysis were a part of just an empirical study. And I am sure that they provide a very poor choices as a "standard" data set and "reference" data analysis.

We acknowledge that the lack of individual-based measures is a limitation of our dataset, and have attempted to make this apparent in the text. However, while there is something to be gained from analyzing individual and dyadic behaviour in the same model, dyadic data tends to be clearly distinguishable from individual data, and provide unique information (Moore et al., 2013). Accounting for non-independence is not always possible if one wishes to spotlight dyadic interactions, as was the case in the published study which produced these data. Moreover, support for multilevel exploratory factor analysis (or PCA) is minimal at best, and the approaches have not been widely evaluated as have, for example, the methods used to determine the ideal number of factors. A comparison of single and multilevel factor analyses showed few differences (D'haenens, 2010).

However, given these concerns, we have added a brief discussion of how Structural Equation Modeling could allow one to model hierarchical, non-independent data.

We are unsure of the intended meaning of the last sentence of this comment; we are honestly uncertain what the reviewer is saying.

The conclusions in the statistical respect are trivial (Kaiser's rule has long been discouraged) and in the primatological respect are also far from conclusive and novel.

We respectfully disagree. The majority of the extraction methods examined in our paper have not been used in prior studies. Previous studies have all used Kaiser’s and scree tests, and one study used parallel analysis. While indeed parallel analysis is more appropriate to use than the former two tests, as we state clearly throughout our paper, parallel analysis should be used alongside multiple – and more robust – tests. We consider additional readily available tests, that are complimentary to parallel analysis and are all more robust than Kaiser’s and scree tests, thereby making our paper a novel contribution to the literature.

This paper does provide a possible explanation for the two different views on the factor structure of social relationships. But it is very far from being clear and convincing.

**We did wish to propose a theory for the explanation, but we must acknowledge that at present, there have not been enough studies to properly tackle this question, and we feel that reaching a conclusion now would be premature. Rather, we wished to encourage researchers to use more robust approaches in future studies, so that the field can (hopefully soon) reach a consensus on this question.**

Any alternative models (e.g. oblique factors) have also not been considered.

**In smaller sample sizes, orthogonal rotations are often recommended, for stability and replicability (Budaev, 2010). Moreover, since prior social behavior studies have largely employed orthogonal rotations (e.g. Koski et al., 2012), we needed to use an orthogonal rotation for the sake of comparative interpretability.**

The generality is also rather week because of the limitations of the study (small sample size, single population, limited situations, scoring and analysis).

**On lines XXX of the revised manuscript, we address issues raised regarding whether or not the results of the present analysis are robust due to sampling inadequacy by stating that “seventy-three dyads and 10 behavioral measures were entered into each PCA, which meets previous recommendations for having a fixed ratio of at least 5 between the sample size and number of variables (Gorsuch 1983).” While indeed we conducted this study on a single population, we state on lines XXX of our revised paper that we are using data on capuchin monkeys as a case study to highlight our more general commentary on extraction methods used in previous studies of social relationship quality in animals. Again, this approach to studying animal social relationships is still relatively new, and thus far has been applied mainly to primates (and one study of corvids). These kinds of analyses are equally useful to studies of other taxonomic groups, but to our knowledge such papers have not yet been published. As a commentary paper, we do not aim to conduct a meta-analysis; instead, we aim to bring these issues directly into the “spotlight” using a single, but clear, example to illustrate our points.**

Finally, the results of the authors (although of course the results are really narrow, single sample with questionable statistical qualities, see below) indicated that all the "tested" methods do agree together while disagreeing with the root-one criterion. Why they suggest that "no single method should be relied upon when deciding how many components/factors to extract" (lns 410-411)?

**A vast body of research (e.g. Budaev, 2010; Ruscio & Roche, 2012) suggests that using a single method to determine how many factors to extract is a risky approach. Seldom will *all* methods agree; we have added material on this issue in the Discussion.**

2. Specific problems

ln. 141 That factor should have complexity of one (variable loading only on a single factor) is reasonable, but strictly speaking, only in the standard psychometric model (items measuring a single concept of thing). In the general sense, especially in animal (social behaviour and other) studies, this is not necessarily true. For example, the same behavioural measure can typically be affected by several psychological or motivational mechanisms, results from adaptive trade-offs between different things etc.

**We agree that complexity is a limited measure; the example the reviewer states is one where a more complicated model might be used. We have added some clarification at this point in the manuscript, and as discussed below, additional material about method choice and SEM.**

ln. 212. N=18 is rather small. It makes low generality (replicability?) for reference data. And it may also be too small for sensible PCA (although maybe not).

**Regarding sample size, on lines 282-284 of our original manuscript, we state “Seventy-three dyads and 10 behavioral measures were entered into each PCA, which meets previous recommendations for having a fixed ratio of at least 5 between the sample size and number of variables (Gorsuch 1983).” In other words, while indeed N=18 in our study, the ratio of items (behavioural measures) to variables (dyads) is well within the recommended level for PCA. Regarding the generality/replicability of our results, as stated on lines 284-286 of our original manuscript, “components with high loadings (i.e. |0.7|) and/or those with four or more loadings greater than |0.4| were considered robust (Guadagnoli and Velicer 1988).” Our components all contained multiple high loadings (>0.7), which highlights the stability of each component. We would also like to again emphasize that the point of our paper is to provide a general commentary using an example of real data to aid with the visualization of our paper’s main point. In other words, questioning the “replicability” of the data specifically used in our paper does not by any means impact the main message of the paper: that researchers using PCA to study social relationship structure must avoid Kaiser’s and scree tests from here on out, and instead turn to using the more robust approaches highlighted in our paper due to the pros/cons we address within our introduction and discussion.**

Lns. 272+. Why such low loadings as 0.4 were chosen as salient? Does this reflect potential statistical significance, confidence interval or any other consideration? Although, in this case quite few such low loadings occurred. Perhaps in studies with small N and wrong model (pseudoreplication) a more sensible approach is to be much more conservative and only take high loadings as salient.

**Using a 0.4 cutoff has been recommended for interpretation in studies of sociality, regardless of sample size (Stevens, 2002). Moreover, 0.4 is a standard in many animal behavior studies (Koski et al., 2012; Weiss et al., 2006; Konecna et al., 2012). We realize that this may be seen as too liberal by some, which is why we required that a factor be characterized by at least 4 loadings of 0.4 or greater, if not characterized by high loadings (> 0.7) in order to be considered robust (Guadagnoli and Velicer 1988).**

**We believe this to be reasonable general advice, but in our own data and extracted solutions, most ratings were quite strong, and ambiguity was minimal, as mentioned in the above comment. Only one primary salient loading was less than 0.5, an alternative cutoff suggested by Budaev (2010).**

ln. 285. Considering variables with many low loadings as salient is dangerous especially with small sample size. Such cases may occur if the model is grossly misspecified or if the variable is not much correlated with any other ones (low R-square).

**As mentioned above, there was only 1 cases where a primary loading did not meet an alternative criteria of 0.5 set by Budaev (2010). The other variable loadings on this component were quite strong. We also mention in a comment below that we did examine the correlation matrix before extraction, and found no variables which appeared to be out of place (i.e. uncorrelated with other measures) among the data.**

ln. 320. This statement ("Correlations between these two components were only weakly correlated (r=-0.072)") is confusing and unclear. Did you calculate any factor comparison coefficients (comparing between two independent factor solutions), like congruence, Kaiser-Hunke-Bianchini or any other?

**The correlation in question is between the two components, within the 2-component structure. We have adjusted the text, which will hopefully make this clearer.**

**Regarding comparison coefficients, we did compute congruence, but did not report them as we thought the results were unclear and less informative than our calculations of omega-hierarchical. The congruence coefficients could be included if it is strongly desired, but we do not believe they would elucidate any particular aspect of the comparison.**

Lns. 268, 369-370. Why the Varimax rotation was chosen? Do the authors have any rationale for computing orthogonal factors? Because of the large differences between the two- and three-factor solutions and a dominance of the first factor I suspect that an oblique solution may be a better model in such a case.

**As mentioned above, in smaller sample sizes, orthogonal rotations are often recommended, for stability and replicability. Since prior social behavior studies have largely employed orthogonal rotations (e.g. Koski et al., 2012), comparative interpretability was a primary concern, hence our use of Varimax rotation.**

Also, did the authors check if all the variables are correlated with the rest (e.g. computed R-square for each individual measure)?

**We did check for large inter-item correlations in the correlation matrix. The highest value was 0.75, with many others above 0.5. All items had a correlation of at least 0.48 with at least one other item. Overall, the correlation matrix appeared to be very reasonably factorable.**

Did they check any factor analysis quality indices like Kaiser-Meyer-Olkin measure of sampling adequacy?

**Yes, but we did not initially report the values. Our overall KMO was 0.7, and this has been noted in the manuscript.**

Were there clearly poor measures that would distort the whole structure (e.g. with low individual KMO?).

**All individual measures were above 0.5 except Conflict Symmetry (0.49), but in both 2 & 3 component structures there was no ambiguity about the loading of Conflict Symmetry with Conflict (which had the second lowest individual KMO of 0.52).**

In my opinion, if the authors aim to provide a sensible "reference" analysis, they should have sensible answers to the basic questions like are the data appropriate to PCA? Orthogonal vs. oblique factors etc.

**We appreciate the reviewers concern for the integrity of this aspect of the research. We have added a section to the results to incorporate these answers.**

Reviewer #2:

This paper deals with extraction methods of components of animal social relationships. The authors compared six extraction methods and assessed how each method affects the results and their interpretations using data of capuchin monkeys. They did so to make recommendations about the use of data reduction techniques for comparative studies of animal social relationships. The paper can make an important contribution, and I agree with the authors that differences between the results of previous studies may reflect instability in the PCA solutions derived using Kaiser's criterion. However, I have a number of concerns that I listed below.

**We thank the reviewer for recognizing the importance of our commentary paper, and have addressed their concerns below.**

-       32-34: The authors stated that "Before performing a data reduction analysis, one must first determine how many factors or components to extract from the analysis", but at this stage it is unclear why this has to be done. Although this becomes clear in the course of the paper, it would be important that the authors clarify it from the beginning.

**We thank the reviewer for pointing this out, and have made the necessary amendments on lines XXX of the revised manuscript.**

-       77-80: The examples of social interactions in the parentheses are not very useful without further explanation. The authors should make a case for listing grooming, tolerance at feeding sites and rates of conflict of conflict as examples of the three listed relationship dimensions.

**We thank the reviewer for pointing this out, and have removed this list of behaviours in the revised manuscript since it is not pertinent to this section of the paper. We have also provided further amendments to this paragraph on lines XXX of the revised manuscript, which we hope will help to avoid any confusion.**

-       82-90: I feel that the argument and the conclusion are misleading. The purpose of using data reduction techniques is exactly the opposite of using a single interaction as a proxy of a dimension. In fact, it is the combination of various interactions that lead to identify the dimension. In addition, the proximity loading varies greatly across studies and is not particularly high in some of them (e.g. Majolo et al. 2010). If the authors want to hold this argument and conclusion, they need to make a stronger case, supported by references and using a logic that is compatible with the approach of data reduction techniques. Otherwise, I strongly recommend the authors to remove this paragraph, especially as this topic is not addressed in the rest of the paper.

**We thank the reviewer for pointing this out, and have removed it in the revised manuscript.**

-       130-132: It is important that the authors explain to the reader how "success" and "correct number of components" were established. Please give detail about which "correct" method was used in Ruscio and Roche (2012) to establish "success" and "correct number of components”?

**We have amended this section to explain the simulations and success rates.**

-        135-188: This section is clear, but it is general. It would be important to conclude with a paragraph arguing which methods are particularly important for extracting relationship dimensions.

We have added a paragraph to this effect.

-       193-196: As we do not live in an ideal world, what can researchers do if they do not find this "ideal result"? What about if the multiple tests do not agree on the same number of components/factors to extract? What solution do the authors propose? Please give more explanation. Simply referring to Everett (1988) in the Discussion is not sufficient for a paper that aims to provide guidelines to extract relationship dimensions.

We have added a section to the end of the paper to discuss this very real issue.

Methods: As the same data were used in Morton et al.'s (2015) PCA, the authors need to clearly explain how the analyses presented here are different from those in Morton et al. (2015). Also, some of the most relevant methods (Kaiser's and PA) were already compared by Stevens et al. (2015). Please acknowledge it more explicitly and explain how your study adds on what Stevens et al. did (by the way they also had the shortcoming of combining dyads from different groups: see below).

**As stated on lines XXX of the revised manuscript, to encourage the future use of more robust extraction techniques, we use data from a previous study of brown capuchin monkeys (*Sapajus sp.*) by Morton et al. (2015) to compare how each of the six extraction methods (described in our introduction) can differentially affect the results, and thus interpretation, of social relationship structure. In other words, we extend the usage of these data by performing XX additional analyses not previously reported in the Morton et al. (2015) paper. We also use the data to provide instructions on how to perform each of the automated extraction tests, and provide recommendations about their use. On lines XXX we explain how our paper extends upon what Stevens et al. did.**

* Table 1: I have several concerns about the behavioral measures.

NOTE: to what extent do our methods match previous studies? Find out – make that the main argument here.

* + 1) Why the first measure in Table 1 is called "Avoid/Stay Symmetry" when only approaches are used in its calculation? How can the authors infer "avoid" from approaches?
  + 2) The measure "Coalitions" should control for the number of opportunities (i.e. A or B being in an aggressive interaction) as done in previous cited studies.
  + 3) Better to use "Aggression" and not "Conflict" as the label for the third and forth measures as conflict may include more than aggressive interactions.

**Unfortunately we are unable to change “conflict” to “aggression” because Table 1 was borrowed with permission from a paper that’s already published, and the authors of this paper used “conflict” as their label for that behavioural measure.**

* + 4) As the measures are dyadic, how did the author decide which member of the dyad was A and B as this is relevant for the four measures dealing with symmetry (e.g. conflict symmetry)? A different index should be used to standardize the four measures regardless who is A and B.
  + 5) The calculation of food sharing is fine only if sharing opportunities are the same for all dyads. Is this possible? Shouldn't the number of times a food resource was eaten by either dyad member be used as the number of sharing opportunities for each dyad?

-       Analysis: Isn't it odd to enter data from two groups in the same PCA? The research question is supposed to be the characterization of social relationships between group members. How combining data from two groups can help in this endeavor? It seems likely that it will mix things up. As the authors aim to give guidelines for the extraction of reliable dimensions of social relationships, I would strongly recommend not combining data from the two groups. If the authors have a good argument to do so, I would still recommend the authors to run also the analyses with data from each group separately because this is the situation encountered in most studies and combing dyads from two groups could create an artificial situation that could affect the validity/generality of the results.

**The purpose of this study is to not assess relationship structure “between groups”, but to assess the structure of this particular “population” of animals, which, by shear availability, happens to consist of two groups. Splitting the data according to social group would reduce stability in the structures and given the already small sample size, could make the solutions uninterpretable.**

**Nesting the groups is possible with more general Structural Equation Modeling, as touched on in Budaev (2010), but it would require introducing a new methodology, and the method is not well suited for exploratory analyses such as these. For these reason and those stated above, we feel that a nested reanalysis is beyond the scope of this paper, though we have added some information about SEM and non-independence of measurements.**

-       276: Please explain in detail how the random eigenvalues for PA were generated in this study and how other researchers can do so.

The random eigenvalues were automatically generated by the fa.parallel function in R, which can be used for any suitable, factorable dataset. We have amended the text to include these details.

-       279-282: In a paper which aims to give guidelines for the extraction of reliable dimensions of social relationships, I would expect more solid indications about the values of component loadings to use as "salient" or "robust". What the authors report seems rather arbitrary. Please provide a stronger case for using 0.4 and 0.7 as "salient" or "robust", respectively, or provide a method to calculate reliably such critical values.

**Unfortunately, the critical values will always be, in a sense arbitrary. Simulations have not yielded any crystal clear criteria due to the wide range of sample sizes and types of data which researchers have used. The criteria are thus the result of discussion among theoreticians (Guadagnoli and Velicer 1988; Stevens, 2002; Budaev, 2010); these criteria are not universally agreed upon, but we believe them to be reasonably conservative given the particular subject area we wish to comment on (social behavior), where past precedent (e.g. Koski et al., 2012) has established 0.4 as salient.**

-       286-287: It is unclear what the authors mean with "2-component PCA solution". And why is it relevant here? Please explain better.

**We apologize, this was a mistake and we have removed it from the revised manuscript. We thank the reviewer for pointing this out to us.**

-       288-294: this paragraph is confusing and its goal is unclear. If the authors plan to focus mainly on PCA (as they seemed saying so in this paragraph), they should state it earlier (in the Introduction?) and justify their selection. They can then focus on PCA and simply mention the relevance of their results to studies using FA in the Discussion.

**We thank the reviewer for this recommendation. In our Introduction of the revised manuscript, on lines XXX, we justify our selection of PCA instead of FA. Specifically, parallel analysis is not appropriate for FA, only PCA. So while five of the extraction methods used in our study need not distinguish between factors and components, PA must be adjusted to support FA (Revelle 2015). In other words, to use PA and to compare those results to the other five extraction techniques, it was necessary for us to focus specifically on PCA. In our revised Discussion, we revisit this issue of PCA/FA comparability on lines XXX.**

-       Figure 2: Some of the text in the figure legend is incongruent with what appear in the figure. Also the two dashed lines cannot be distinguished. Please use two different types of lines and revise the figure legend.

**One of the lines has been altered in order to be more easily distinguishable from the other, and the legend has been fixed.**

-       Shouldn't also a line for the 95th percentile of the PA data be shown in Figure 2 to help the reader in following what the authors stated in line 307?

**Error bars have been added to Figure 2 to assist the reader.**

**Figures 1 & 2 have been taken as directly as possible from the output of the R functions nFactors and parallel, respectively. As such, we do not wish to alter them except where it is absolutely necessary, for we believe it would create greater confusion when readers attempt to use our guidelines, and find the results inconsistent with the approach we recommend.**

-       299: Isn't this statement a bit arbitrary? Looking at Figure 2, it is not clear that the scree plot suggests only 2 components. Why not 3 or 4?

**The scree plot is not ambiguous, as only components which are above the resampled data line (and outside the error bars) should be retained for extraction. Since we are using PCA, we examined the PC line and found that PA recommended the extraction of only 2 components. The text has been amended to explain this in more detail.**

**Figures 1 & 2 have been taken as directly as possible from the output of the R functions nFactors and parallel, respectively. As such, we do not wish to alter them except where it is absolutely necessary, for we believe it would create greater confusion when readers attempt to use our guidelines, and find the results inconsistent with the approach we recommend.**

-       318-320 vs. 334-336: The authors plan to propose guidelines and reduce arbitrary decisions, thus they need to be very careful in their terminology. How can Component 2 in the first PCA and Component 3 in the second PCA have the same variables with high loadings (conflict and conflict symmetry), but the interpretation of the meaning of the component differs between the two cases: simply "agonistic behaviors" for Component 2 and "behaviors indicating a lack of stability or predictability in the relationship" for Component 3? Apart from this inconsistency, it is unclear why high loadings of conflict and conflict symmetry should be interpreted as lack of stability or predictability. High rates of aggressive interactions from both partners seem a highly predictable aspect of a social relationship.

**We agree and thank the reviewer very much for making these valid points. Because Kaiser’s criterion was the only method that recommended the 3-component solution, and as we discuss throughout our paper this method is highly inappropriate for interpreting the meaningfulness of one’s results due to it creating structural instability within the PCA solution, we opted to delete all mention of how those results reflect Cords and Aureli (2000) since there was little evidence to support the notion that the way in which items loaded onto each of the three components were biologically meaningful. Instead, we provide the labels C1, C2, and C3 to those components for the purpose of referencing them throughout our paper. Also, as stated on lines XXX of the revised manuscript, while structural comparisons between our 3-component solution and those reported in other studies *may* be socially meaningful, they are more likely to reflect the structural instability of this solution brought on by over-extraction and spreading items “too thin” across components. This would explain, for instance, why the first and second components contained overlapping items with moderately high loadings (Table 2); the internal consistencies of both components suggested that a more general latent variable (e.g. a combination of two of the extracted constructs) could underlie the overall solution.**

-       320-321: please revise this sentence as a correlation cannot be correlated.

**We thank the reviewer for pointing this out, and have fixed this on lines XXX of the revised manuscript.**

-       337-338: Please provide information about the procedure to assess the internal consistency of the solutions and what high/low values of ωh mean in the Methods. Otherwise, the text in lines 338-344 is not understandable.

**We have added text to the methods section to explain the meaning and interpretability of omega-hierarchical.**

-        347-351: Yes, the 3-component solution has the same number of components as the ones proposed by Cords and Aureli (2000), but the rest of the comparison is not as straightforward as the authors believe. 1) Their Components 1 and 2 have Coalitions and Social Foraging in common, behaviors that are expected to reflect Value. 2) The authors interpret tolerance at feeding sites as an element of Compatibility, whereas such a tolerance is often view as service provided in exchange of grooming (e.g. Barrett & Henzi) and so it should be viewed as Value. 3) As I argued above, Component 3 should not be labeled as Security given that Conflict and Conflict Symmetry do not really capture unpredictability (please note that the component with high loading of aggression is usually interpreted as [In]compatibility in previous studies). See Fraser and Bugnyar (2010) for an appropriate example of results supporting Cords and Aureli's proposed three components.

**We agree and thank the reviewer very much for making these valid points. Because Kaiser’s criterion was the only method that recommended the 3-component solution, and as we discuss throughout our paper this method is highly inappropriate for interpreting the meaningfulness of one’s results due to it creating structural instability within the PCA solution, we opted to delete all mention of how those results reflect Cords and Aureli (2000) since there was little evidence to support the notion that the way in which items loaded onto each of the three components were biologically meaningful. Instead, we provide the labels C1, C2, and C3 to those components for the purpose of referencing them throughout our paper.**

-       359-361: Please explain how Stevens et al.'s "findings reflect how a scree test …. can sometimes be more liberal than the unbiased parallel analysis". Too little information is provided for the reader to appreciate this inference.

**We apologize, this was a mistake and we have removed the sentence from the revised manuscript. We thank the reviewer for pointing this out to us.**

-       366-378: This part is affected by the misinterpretation of the 3-component solution (see two comments above) and should be revised accordingly. In addition, the authors may want to consider/include in the revision that behavior indexes such as grooming symmetry and other symmetry indexes do not have an apriori link (i.e. they are not obvious predictors) with any of Cords and Aureli's components and are probably entered in various studies' PCA because they are easy to calculate, increasing the number of variables that otherwise may be too low to run a meaningful PCA.

**We thank the reviewer for making these recommendations, and have revised this portion of the manuscript (see lines XXX). In particular, because Kaiser’s criterion was the only method that recommended the 3-component solution, likely leading to structural instability within the PCA solution, we opted to delete all mention of how those results reflect Cords and Aureli (2000) to avoid this issue entirely. We agree that certain behavioural indices may not have an apriori link to Cords and Aureli’s proposed model, but we feel that this is beyond the scope of the main goal of our paper, and rather should be the subject of a different paper that addresses that issue more satisfactorily. These behavioral indices have already been used extensively throughout published relationship studies, so we opt to retain these indices for the sole purpose of comparability to other studies.**

-       Discussion: in various places the authors tend to consider the 3-component solution, often found using the Kaiser's criterion, as an expression of Cords and Aureli's 3-component proposal, but these are issues at two different levels, empirical and theoretical. The number of extracted components should not be the criteria to support Cords and Aureli's 3-component proposal or Hinde's 2-component proposal. It is the loading in the components that should help the interpretation/labeling of the components. Rebecchini et al. and Stevens et al.'s extracted 2 components, but they did not label them simply as the Affiliative and Agonistic dimensions because the variable loadings suggested more meaningful, although arbitrary, interpretations.

**We very much appreciate this comment. In addition to our responses to the reviewer’s last comment, we have adjusted our revised manuscript to accommodate these valid points (see lines XXX).**

-       409-416 and Conclusion: It would be important to have a stronger statement about "what to do". It would be more helpful for readers/researchers if the authors would give a clear "receipt". Otherwise, the take-home message "do a bit of everything hoping for consistency of the results" could have been provided at the beginning of the paper without the tests on the capuchin data. In addition, as Kauser's criterion is the default implementation in IBM SPSS (line 133) and possibly other software, it would be important for the authors to comment on how to implement the proposed methods, especially PA, in IBM SPSS and other statistical software (perhaps in an Appendix).

**We thank the reviewer for these useful comments. On lines XXX of the revised manuscript, we provide details on “what to do” with regards to running these analyses in R. We use R because 1) it’s widely used, 2) the program is free to the public, 3) it is easy to download (for free) and run code for extraction tests not offered (or difficult to find code for) in SPSS.**

**In the revised manuscript, we clarify the notion that researchers must “do a bit of everything hoping for consistency of the results”. Specifically, we state on lines XXX that all methods of course have their drawbacks (Ruscio and Roche 2012). Therefore, no single automated extraction test should be used as the sole method to determine how many components/factors to extract from a data reduction analysis. This is predominately because few datasets yield an immediate and clear solution. Ideally, multiple automated tests should be implemented and compared; if multiple tests agree on the same number of components/factors to extract, then researchers can be confident with their decisions about extraction (refs).**

**We discuss in the revised manuscript what researchers should do in cases where multiple automated methods do not agree on the same number of components. Specifically, on lines XXX we state…**

**Finally, while indeed we could have written this commentary without the use of the capuchin data, we opted to use data to provide a visual illustration of the various extraction methods employed here. We opted to use real data on capuchin monkeys, rather than simulated data, because it was important for us to test the robustness of the extraction methods used in the published study by Morton et al. (2015) using the same dataset. In other words, we have provided evidence that the 2-component solution published by Morton et al. (2015) – which used a scree test and only a single automated extraction test – is indeed the best structure given the item-to-variable ratio. On lines XXX of the revised manuscript, we address issues raised regarding whether or not the results of the present analysis are robust due to sampling inadequacy by stating that “seventy-three dyads and 10 behavioral measures were entered into each PCA, which meets previous recommendations for having a fixed ratio of at least 5 between the sample size and number of variables (Gorsuch 1983).”**

-       The Conclusion does not add much and can be deleted/incorporated in the previous paragraph, while providing a more meaningful guideline for future research (see above).

**We appreciate the reviewer’s opinion, but we would like to retain the conclusion section because we believe it to be a nice, abridged summary of the key points of our paper. We have shorted the conclusion in order to tighten the overall message up more succinctly. However, if the reviewer still feels strongly about removing the conclusion section, we are happy to do so.**

* 419-420: The authors wrote "The current example should serve as a cautionary note to researchers wishing to use data reduction analyses to study social relationship structure in animals", but as pointed out above there are several shortcomings in this example, starting from the behavioral measures and the combining of dyads from two groups (see above).

**In general, we feel that while our data have shortcomings, all datasets do, and we believe that the issues in this dataset are not unlike those which other researchers have and will face. However, we have addressed these concerns as best we are able in previous reviewer responses (see above) and on lines XXX of the revised manuscript.**

* As any study has some intrinsic problems, it would be important to have a larger sample of examples to verify whether the findings of this study are representative, given that in several previous studies the variable loadings matched very nicely Cords and Aureli's 3 components. Therefore, I encourage the authors to carry out multiple comparisons of the various methods (similar to the one done on the capuchin monkey data) using other datasets. These additional comparisons will be helpful in evaluating the generality of the findings of this study and corroborate them given the potential problems of the current dataset (see my

comments above). The authors may include additional datasets by asking authors of cited articles, perhaps in exchanging of coauthorship.

XXX