**Dear Dr. Ueno,**

**Thank you for your helpful comments about our revision. We have updated the manuscript in line with these comments, and we present our responses below.**

Comments to the author (if any):

<< Overall Comments >>

This study deals with the effect of randomization of the presentation order of test items on the psychometric characteristics of psychological scales. Analyzing real data on psychological scales with multiple methods, the authors found that randomization made little difference to the psychometric properties of the scales.

The manuscript was reviewed by two independent reviewers. Both reviewers are positive as to the subject and approach of this study. However, Reviewer #1 raised some technical/methodological questions, while Reviewer #2 requests clarification of some of your ideas and procedures. Most of the comments from both reviewers sound very reasonable in order for the manuscript to be improved further. Thus, below I compiled a list (1-8) of their major comments (plus my own ones), each of which I request you to address in an appropriate manner in the revision. The revised manuscript will mainly be evaluated for how much it has been improved with respect to these points. In addition, it would be greatly appreciated if you could go through and respond to each specific comment as much as you can.

(The reference to the corresponding reviewer comment is indicated in parentheses in each item.)

1. You need clearer and deeper argument that supports the idea that randomization of item order alleviates order effects (#2[1]). Probably there are different kinds of order effects, and randomization may not be the only way to deal with them depending on the nature and purpose of the problem. In this regard, your literature review seems to cover the relevant topics fairly well but might miss some critical references. For example, I was able to find classic but excellent chapters in the book below (PART IV, chapters 15-17) that may help you better frame your research problems and methodology.

- Schwarz, N., & Sudman, S. (Eds.) (1992). Context effects in social and psychological research. NY: Springer-Verlag.

**Thank you for the reference for this book, as it was very helpful. We have used this information to add to the introduction, hypotheses, and discussion. Mainly, we added how the four-part integration model and meaning-change hypotheses describe how and when context ordering effects may occur, as well as added literature on classic effects of ordering.**

2. You also discuss some methodological issues in the introduction. However, it is surprising that you do not mention the psychometric concept and relevant methodology of "measurement invariance" (MI, or "measurement equivalence"), which seems highly relevant to this study (yes, you use the same term on page 3 but that is in a different context; what I call MI here is truly a psychometric concept). I request you to include the literature of MI and consider using some method(s) for testing/evaluating MI in this study if you think you need to do so based on your review (and even if you do not, please explain why). Below I list two fundamental references for MI for your information.

- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58, 525-543.

- Millsap, R. E. (2011). Statistical approaches to measurement invariance. NY: Routledge.

**To focus on tightening terminology, we used the term equivalence to describe all types of equality in scales, while describing measurement invariance as a psychometric concept. The details on equivalence and invariance were expanded in paragraph 2 of the introduction and added to the data analysis section. Generally, this paper focused on easy ways to implement equivalence testing that any researcher with a basic statistics background could analyze. We did test measurement invariance through MGCFA and added that information in a supplementary document online. If the reviewers think this information would strengthen the paper, we can add it to the document instead.**

3. Please reconsider the method of comparing two sample covariance matrices or provide justification for using the current method (#1(1)).

**Thank you for the recommendation of cortest in the psych package. We have changed these analyses to use that test have updated the results and discussion accordingly.**

4. The descriptions about the Bayes factor are not very accessible (#2[11]), and some of them are even pointed out as inaccurate (#1(3)). This seems to come from the authors' insufficient understanding of the (general) Bayesian methodology. Please consult the suggested reference (Lee & Wagenmakers, 2013) and maybe some others, and make appropriate corrections.

- Lee, M. D., & Wagenmakers, E.-J. (2013). Bayesian cognitive modeling: A practical course. Cambridge, UK: Cambridge University Press.

**We have updated this section using the book recommended, along with other related sources from Wagenmakers and colleagues.**

5. Please provide justification for using d\_av (#1(4)). Also, please make your description consistent and clearer in order to avoid possible misunderstanding (#2[10]).

**D\_z is known to overestimate the size of effects, and therefore, we used d\_av as the effect size for dependent t, based on recommendations from Lakens and Cumming. Additionally, d\_av is more comparable to d\_s for independent t, which is also used in the study. The language and references here was edited to reflect this reasoning better.**

6. Please provide more details about the data screening procedure and its rationale (#1(2)).

**This section was expanded to address the reviewers points – and those responses are described below. Mostly, we added more text explaining our choices and process.**

7. Please indicate summary statistics for those scales in each mode, including the score mean, SD, reliability, and standard error of measurement (#2[7]).

**The means, SDs of each scale in each format are provided in Tables 3 and 5. We re-organized the columns based on the reviewer’s comments. The reliability of each scale for each modality was added to the total scores section of the results. We did not include SE in these tables, as they can be calculated using the final sample size listed in Table 1 and the SDs in each table.**

8. The fourth paragraph in the discussion section (page 18, L429-) should be rewritten to include more practical concerns and to state the following things more clearly:

- Please include your reaction to the comment #2[13].

- If the randomization makes difference, what should you do? One of my (practical) concerns is that it seems to me that randomization of item order, whenever effective, works for canceling out item reactivity at the \*group\* level, but this has nothing to do with obtaining valid \*individual\* scores. Again, what should you do then? It would be even better if you can include your consideration of which indicator/method implies what.

- If the randomization makes no difference, can you just leave it as is, or ...? (you mention further investigation of sample specificity and that is one of the good points.)

**Given the literature on how participants create context through the four-stage response process and meaning-change hypothesis, it would be important to examine how randomization may affect responses. Either way, it is practically important to know if I randomize my scale, and another research does not, are these comparable sets? We have added to the discussion to comment on subject matter experts and individual versus group considerations.**

<< Specific Comments >>

**Our referencing software included all first initials for authors whose last name was represented multiple times (i.e., Brown, Buchanan) based on APA style requirements. We have since figured out how to correct this behavior, so that these initials do not appear to correspond to normal APA use.**

P3, L34: "T. Buchanan" shall be "Buchanan."

**Fixed.**

P3, L37-38: "T. Brown" shall be "Brown."

**Fixed.**

P4, L45-46: "the nature the online world" shall be "the nature of the online world."

**Fixed.**

P4, L60: "T. Buchanan" shall be "Buchanan."

**Fixed.**

P4, L64: "A. Weigold" shall be "Weigold."

**Fixed.**

P4, L70: "C. Cook" shall be "Cook."

**Fixed.**

P6, L103-104: "Factor loadings represent the correlation between each item and the overall latent variable" - This is not very correct. This holds only when the one-factor model is used or the factors are uncorrelated with each other. The correlations between item scores and factors are sometimes termed "factor structure" (also see the related comment of Reviewer #2 below).

**We have rephrased this line based on this comment.**

P7, L138: "item loadings" sounds strange to me; what you mean by "item loadings to their latent variable" seems to be "factor loadings."

**We have rephrased item loadings to factor loadings for more consistency in terminology.**

P8, L164: "question loadings" - the comment right above also applies here.

**We have rephrased question loadings to factor loadings for more consistency in terminology.**

P8, L173: "by summing questions" should be "by summing item scores."

**Fixed.**

P9, L190-191: "Scales were randomized across participants" shall be "The order of administration of the two scales was randomized across participants" ?

**We have rephrased this sentence to reflect that the order of administration of the two scales was randomized across participants for both groups.**

P10, L214: "if the correlation matrices are different" shall be "if the covariance matrices are different" ?

**This section was updated as part of the change in analysis recommended by another reviewer.**

P10, L216: "We then calculated an exploratory factor analysis" shall be "We then conducted an exploratory factor analysis."

**Fixed.**

P11, L244: "E. M. Buchanan" shall be "Buchanan."

**Fixed.**

P13, L293: Comrey and Lee (1992) is not in the reference list. Please add.

**Fixed.**

P14, L322: You state "only 3 values were significantly different" but only two cases are shown here?

**This analysis has since changed based on the recommendations of another reviewer.**

P17, L397-398: Do the terms "scale invariance" and "survey equivalence" here mean the same thing as "measurement invariance?" Please try to use the same term consistently in order to point to the same thing.

**To focus on tightening terminology, we used the term equivalence to describe all types of equality in scales, while describing measurement invariance as a psychometric concept. The details on equivalence and invariance were expanded in paragraph 2 of the introduction and added to the data analysis section.**

P17, L403-405: "Decreased variance typically results in decreased measurement error" - This is not necessarily true. A general notion in the classical test theory is rather opposite; if the true score variance becomes larger, so does the observed score variance. Usually, items (or tests) with smaller variance are less discriminating among individuals (e.g., Crocker & Algina, 1986), leading to decreased reliability. Accordingly, your following statement "randomization has the potential to decrease measurement error" cannot be guaranteed. Please reconsider these statements, or provide additional evidence that the measurement error has actually decreased by randomization.

- Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Harcourt Brace Jovanovich.

**The reviewer has an excellent point. As part of the analysis update suggested by another reviewer, this section has been changed completely.**

P19, L451: "T. Brown" shall be "Brown."

**Fixed.**

P20, L462: Only the first word of the book title should be capitalized (Brown, 2006).

**Fixed.**

P26, L621: Only the first word of the book title should be capitalized (Tourangeau et al., 1999).

**Fixed.**

Tables 1 and 2: It would be better to exchange Columns 5 (SD-R) and 6 (M-NR) to make it easy to compare the corresponding indices between the random and non-random conditions.

**Updated.**

[Comments from Reviewers]

Reviewer #1:

This paper examined whether (and if so, where) there are differences in scale relationships for randomized and nonrandomized versions of psychological scales. The authors found that the psychometric properties of a scale are not much affected by item randomization. My general impression is that the authors investigated an important research question, and that their findings will add to the existing body of knowledge in psychometrics and behaviormetrics. This article is generally well-written, with informative contents and clear explanations. That being said, I think that the study has several notable issues that need to be addressed.

(1) Comparison of covariance (or correlation) matrices

The authors compared the covariance matrices between the randomized and non-randomized forms using RMSE, with "a criterion of <.06 for good fit, .06-.08 for acceptable fit, and >.10 for bad fit was used (Hu & Bentler, 1999)" (p.10). I am suspicious of this method and the cutoff values for several reasons.

First, I could not find reference to the above cutoff values in Hu and Bentler's original paper (1999). I believe that the "RMSE" used in this study corresponds to the "SRMR" (standardized root mean square residual) used by Hu and Bentler (1999), but I could not, nevertheless, find an explicit statement for these cutoff values. Note that "RMSE" is not mentioned in Hu and Bentler (1999), and "RMSEA" appears to have a different definition. Second, the research purpose is different. Hu and Bentler (1999) investigated cutoff criteria for evaluating the fit of a \*single\* model of interest (sometimes against the null or full models), whereas the current study used such criteria for comparing \*two\* sample covariance matrices. Third, there are more common ways of comparing two correlation matrices. For example, the cortest.mat function in the psych package of R provides a chi-square test for determining whether a pair of correlation matrices are equal, which might be more

appropriate for the current purposes. A Bayesian approach would also be possible.

Accordingly, I would like the authors to consider a better method of comparing covariance (or correlation) matrices. I think this point is important because changing the method might lead to a change in the conclusion. Particularly, considering the fact that only a few elements of the covariance matrices were significantly different (p.14), and that, according to the NHST framework, 5% of the results will theoretically be significant even when there is no difference, I think it might be better to consider that the covariance matrices generally do not differ.

**Thank you for the thoughtful response in this section. We switched these analyses to using cortest.mat, as it seemed much more appropriate. This analysis allowed us to use more traditional measurements of significance and z-score comparisons. These sections of the results and discussion were updated accordingly.**

(2) Data screening

I think the "Data Screening" section (p.14) in its current form is insufficient, and that further justifications and clarifications are required. First, I am not sure why the authors use different methods for dealing with the data above and below 5% missingness; it would be necessary to justify this handling.

**This percentage is based on recommendations from Tabachnick and Fidell’s book chapter on data screening. We added that information to this section.**

Second, the authors used the mice package of R to impute all observations with <5% missingness. I think that the authors intended to use multiple imputation because the mice package is used for that purpose; however, no further description of how the authors conducted the multiple imputation is provided in the manuscript. For multiple imputation, one must (1) generate multiple imputations, (2) analyze the imputed data, and (3) pool the analysis results, none of which is described in the current manuscript.

**We added more information about mice and our imputation to the data screening section.**

Third, the number of imputed observations is not described in the manuscript or in Table 1.

**We included the number of observations imputed in the data screening section. We added the final N to the first table, so it logically subtracts from original N to final N. We thought that adding imputation numbers to the table might be confusing as they are part of the original total and final sample, so they are simply stated in the paragraph.**

I further think that it would better for the authors to clarify the final sample size in Table 1.

**We have added this information to Table 1.**

Fourth, the data exclusion based on the Mahalanobis distance needs further explanation and justification.

**We added more information to this section.**

Fifth, because the number of excluded observations is relatively large in Table 1, the authors might discuss why they believe it is valid to exclude such a considerable number of observations from further analysis.

**We added information about why we excluded these participants in the data screening section.**

(3) Bayesian methodology

I am a Bayesian (although I have respect for frequentist methods), and from my point of view, the current description on the Bayesian methodology and Bayes factors (p. 12 as well as several other places) is not very accurate. I would like the authors to consult the Bayesian statistics literature and thoroughly rewrite their descriptions on the Bayesian methods. Some of my concerns are as follows:

- It seems to me that the authors confuse the prior distribution on the parameter and on the hypothesis. The "recommended default priors" (p.12) used in the BayesFactor package is the prior on the \*parameter\*, not on the hypothesis; however, some of the authors' descriptions (e.g., "Prior distributions are our estimation of the likelihood of our hypothesis before the data was collected" (p.12)) read as if the authors meant the prior on the \*hypothesis\*. One of the attractive features of using Bayes factors is that no matter how the authors set the prior on the two hypotheses, it does not affect the resultant Bayes factors. Of course, the priors on the parameter do affect the resultant Bayes factors, which is why we need the "default priors."

- "Traditional NHST focuses on the likelihood of the data, given the null hypothesis is true, and Bayesian analysis instead posits the likelihood of a hypothesis given the data." (p.12) This statement may not be very appropriate because the Bayes factor is different from the probability of a hypothesis. Probabilities of hypotheses were not calculated in this paper.

- Bayes Factors are ... a ratio of the likelihood of two models" (p.12). This is not correct; in fact, it is a ratio of the \*marginal likelihood\*, not the \*likelihood\*, of two models. These are two different concepts.

- Descriptions like "the BF for this test was 0.24 ±0.02%" (p.15; similar descriptions are found in other parts of the manuscript as well) may be inadequate because the Bayes factor is a ratio. All ratios are not symmetric around the point estimate. Therefore, I think the interval estimates of Bayes factors should essentially be asymmetric, and thus cannot be represented by the "±" sign.

I think an appropriate study that the authors can consult would be Lee & Wagenmakers (2013), especially Chapter 7.

- Lee, M. D., & Wagenmakers, E.-J. 2013. Bayesian cognitive modeling: A practical course. Cambridge, UK: Cambridge University Press.

**We have rewritten this section after consulting the book mentioned and a Bayesian friend. This section should now more accurately reflect Bayes and BF, along with more references for interested readers. We would still like this section to be approachable for traditional NHST researchers and gladly would take other recommendations if necessary. The ± information has been removed, as it was a misunderstanding of the BF output.**

(4) Effect size measure

The authors report the d\_av, which ignores the correlation between two dependent measurements, and thus should essentially be biased unless the dependency between the two measurements is ignorable. The authors state that "This effect size is less biased than the traditional d\_z formula, wherein mean differences are divided by the standard deviation of the difference scores (Lakens, 2013)." (p.11). However, based on my own reading, Lakens (2013) did not generally prove or even state that d\_av is less biased than is d\_z. Note that the \*bias\* of an estimator is a statistical property that can be mathematically derived. While I do not oppose the use of d\_av in this study, please provide some valid justifications for it.

**We have revised this section to indicate that we were referencing the overestimation of d\_z. d\_rm is also a useful measure of effect size for repeated measures, as it does control for the correlation between levels. However, we used d\_av additionally as a measure that would be comparable to d\_s, which is also included in the study. We detailed this information in the data analysis section.**

Minor points:

- Reverse coding is only mentioned for the LPQ, but not for the PIL. Please explicitly mention that there are no reverse items in the PIL, if this is the case.

**The PIL does not technically have reverse coding, but the end points are switched high to low on some items (and therefore, practically has reverse coding depending on how the data is “entered”). We added that information to the materials section.**

- "NHST has also been criticized for an inability to test the null hypothesis, ...." (p.12). I think this statement is incorrect. In fact, the NHST \*does\* test the null hypothesis, and thus can reject it. However, the framework of the NHST cannot provide evidence \*in favor of\* the null hypothesis.

**This was phrased poorly, as you mention, and has been reworked to clarify.**

- The Open Science Framework website to which the authors refer on p. 13 (<https://osf.io/6qxdn/>) provides not only the "graphics created from this package" (p.13) but also the raw data and other useful materials. I support such an open practice, and would like the authors to explicitly state that that the raw data and other useful materials are provided at the OSF website.

**Added this information in a couple of sections.**

- Typographical errors:

-- withing a 5% -> within a 5% (p.17)

-- Frankl's concept ofnoogenic neurosis -> Frankl's concept of noogenic neurosis (p.21)

**Fixed.**

Reviewer #2:

(Paragraph numbers were added by the Associate Editor. Otherwise the text is intact.)

This is an interesting study with the potential to be informative to other researchers. The study is fairly clearly described and the use of multiple indicators of effects was appreciated. There are several aspects of the study description that could be improved to make the results clearer and more accessible.

[1] The argument regarding the use of randomization of items is not clearly stated. Why would one want to introduce the possibility of variation in item response and scores by randomizing item order? In achievement testing, in many cases, we see significant variation in item difficulty based on item position (item position effects). Not only is this the result of context effects (the effects of neighboring items) but also simple order effects (an item tends to be easier when it appears at the beginning of the form rather than the end of the form). In other cases, randomizing items supports item and test security - to prevent cheating by test takers sitting in close proximity. But in your case, it seems you are arguing that if there are context effects, we should minimize those effects by randomizing item order. This argument could be more directly and clearly described.

**We used the references provided by the editor to improve the introduction and discussion as to this point. Additionally, as explored in the discussion, it would be useful to know the effects of randomizing item context – which would allow different modalities of presentation to be compared knowing they are equivalent (minus other experimental factors).**

[2] On page 6, there are a number of statements that could be clarified. First, you argue that EFA provides factor loadings and overall model fit - however, if it's truly EFA, there are no commonly agreed upon indicators of model fit. These are more likely found with CFA. Either way, your study does not examine overall model fit for the EFA results.

**In the introduction, we have briefly summarized how a scale might be designed and analyzed, as context for scale development. Model fit is discussed in most texts and how-tos for EFA, as exploring fit indices like RMSEA or CFI. While there are many fit indices, and people argue about “gold standards” for them, it appears that they are used in the literature to describe fit between the data and the model. We left this section in, but would be happy to revise if the reviewer can clarify the issue with this section.**

[3] Next, you describe factor loadings as the correlation between each item and the overall latent variable. This is not the case. The factor loadings are correlations between item scores and the factor score (which is not the latent variable).

**This section was corrected.**

[4] Also on this page you parenthetically refer to "relation to other phenomena" regarding construct validity. First, your use of the term "construct validity" is consistent with old notions of validity (as types) rather than validity as a unitary concept. But moreover, there are many forms of evidence that provide construct-related evidence of validity. And regarding associations with other phenomena, such evidence also constitutes criterion-related validity evidence. Anyway, it's strange that you describe score reliability and validity evidence since your study provides neither.

**We rephrased this slightly to talk about validity as a whole (rather than just construct, as there are lots of types, as you mention). We also added reliability measurements to the study.**

**Potentially, we could shorten this paragraph, but we included it as an overview of how scales are developed as a framework for how analyses are traditionally performed to develop the scale focusing on the items, and then the items are often ignored as a unit of analysis. But, as the next new section argues, items and their context are important for understanding how participants answer the scale.**

[5] Then, it is here on page 6 where you argue that "the theory" needs information regarding the psychometric invariance of item randomization.

**This section has been updated, which hopefully resolves some of the reviewer comments.**

[6] A minor note: On page 8, and elsewhere, you describe the 7-point "Likert type response format" and then argue that "however, each item has different anchoring points" as thought that was not a feature of "Likert type" response formats. First, I'm not sure what makes these 7-point rating-scale items Likert-type items, but in Likert's work, his items were also tailored so that the response options matched the content of the item stem. However, the main issue is that Likert did not introduce rating scales and did not introduce a specific response format. His dissertation/monograph was important as it simplified the scoring of attitude measures - but referring to rating-scale items as Likert type is a misnomer.

**We have edited this section to reflect this point.**

[7] Also on page 8, you refer to "the reliability for the scale" as generally high (and similarly elsewhere in the manuscript). Reliability is not a characteristic of the scale/measure/test. It is a characteristic of the scores resulting from administration of a measurement tool. Reliabilities are sample specific and vary significantly as a result of changes in sample composition (particularly sample variance). The practice in educational and psychological research is to report reliabilities of the scores in hand - the reliabilities of scores used in the particular study. You should report reliabilities for each measure/administration mode combination.

**We have added the reliabilities to the total scores section and edited the text around reliabilities to convey this point.**

[8] On page 10, there is a statement that appears to be contradictory: that factor loadings were examined in case "participant interpretation of the item changed the relationship to the latent variable. However, we did not predict if values would change, as latent trait measurement should be consistent." If factor loadings do change, and this indicates that participant interpretations shifted so that item associations with factor scores also shifted, then those factor scores are likely representing a shift in the latent variable itself. As items become more or less associated with the factor score, our interpretation of what we are measuring also changes - as we may not be measuring the same thing.

**We tried to reword this section to be clearer that we weren’t sure if the factor loadings would change or not.**

[9] Also on page 10, you indicate that you conducted an oblique rotation in the EFA. Rotation is not possible if you extracted one factor.

**Fixed.**

[10] I appreciated the inclusion of the standardized effect sizes to do comparisons. However, I wasn't familiar with the specific formula you used. You take an average of SDs, where it is more statistically consistent to average the variances and take the square root for the standardized difference. At any rate, I was really confused by your description at the bottom of page 11, since you describe what you're not doing at the same time as what you are doing. So I don't even know if this is the d formula you use in your analyses, since you also state that you prefer to use the SD of the difference - which doesn't seem to be what is included in the formula.

**This section was reworked to help clarify this point, in addition to a separate reviewer point. In short, d\_z can overestimate an effect, and d\_av may represent the effects better. In addition, d\_av is comparable to between subjects formulas d\_s, which are also used in this study. When using repeated measures, we like to provide the formula and description of our effect, as reporting it only as *d* is ambiguous.**

[11] I don't know that the BF results provide any additional support for your argument/analyses. This is a complicated section in your description (p. 12). You don't provide any information regarding how the choice of priors relates to your data or measures (particularly the Jeffreys prior). Why is this relevant? What does this prior have to do with your research design or the data distributions?

**We reworked this section based on other reviewer comments. We wanted to be able to include an analysis that would show support for the null in comparison to an alternative model. BF are an excellent way to address that question – while TOST can also answer these questions, some may consider it to suffer from the same issues as NHST. This new section should address these issues.**

[12] As a side note, I noticed a great deal of consistency in the factor loadings between random and non-random administration within measure. The correlations between random/non-random factor loadings within measures was .96 for both measures. The correlations of factor loadings between measures were not as high, but they were higher for the random administrations (.80) than for the non-random administrations (.68).

**Interesting!**

[13] I greatly appreciate the cautions provided at the end of the paper, particularly regarding the potential of sample specificity of results. This is a good example of how to investigate the question of random-item administration. But the question should be driven by a strong argument for the need to do so. If there is potential for strong order effects, because of the nature of the items, this might be a consideration. But content experts must review the order of items and consider the potential of item-order effects. And as you argue, empirical evidence of the psychometric equivalence given administration method must be gathered. The need to do random item administration must be clearly articulated first.

EOF

**Thank you for the feedback! We believe the improved introduction and discussion will address some of these comments.**