From: Loïs Fournier (corresponding author), on behalf of all the authors involved (Alexandre Heeren, Stéphanie Baggio, Luke Clark, Antonio Verdejo-García, José C. Perales, and Joël Billieux)

To: Veli-Matti Karhulahti (recommender) and Ivan Ropovik (reviewer)

Dear Veli-Matti Karhulahti and Ivan Ropovik,

We were pleased to receive the decision with respect to the second version of our stage one registered report manuscript, *Development and evaluation of a revised 20-item short version of the UPPS-P Impulsive Behavior Scale* (Manuscript ID: 862), in your communication of November 20, 2024.

We thank you, Veli-Matti Karhulahti, for your willingness to consider a third version of our manuscript.

To facilitate the review of the third version of our manuscript, comments are numbered and presented in table cells, below which are our responses. All revisions to the original version of our manuscript have been highlighted in green in the (.pdf) present document. All revisions to the original version of our manuscript have been highlighted using the “track changes” function in *Microsoft Word* in the (.docx) revised version of our manuscript.

I, Loïs Fournier, act as the corresponding author of the present revised version of our manuscript and remain at your disposal for any further information.

Sincerely,

Loïs Fournier (corresponding author), on behalf of all the authors involved (Alexandre Heeren, Stéphanie Baggio, Luke Clark, Antonio Verdejo-García, José C. Perales, and Joël Billieux)

Loïs Fournier, M.Sc. (lois.fournier@unil.ch)
Institute of Psychology, University of Lausanne, Lausanne, Switzerland

From: Loïs Fournier (corresponding author), on behalf of all the authors involved (Alexandre Heeren, Stéphanie Baggio, Luke Clark, Antonio Verdejo-García, José C. Perales, and Joël Billieux)

To: Veli-Matti Karhulahti (recommender)

Comment VMK-00

> Dear Loïs Fournier and colleagues,
>
> Thank you for your patience with a small delay. One of the reviewers returned with further feedback and I think it was well worth the wait. I won't reiterate it here but let you engage with the comments directly – a couple of small technical follow-ups.
>
> The idea of RRs isn't to overly burden authors or reviewers, and the manuscript is close to IPA, so I likely won't invite reviewers for more rounds but will evaluate the final revisions myself (consulting PCI statistics if open questions remain). Therefore, please make last steps in proofreading carefully to minimize changes at Stage 2. As usual, you can contact me pre-submission if specific questions occur. Again, thanks for all detailed responses to this interesting work.
>
> Best of health and wishes,
>
> Veli-Matti Karhulahti

Dear Veli-Matti Karhulahti,

Thank you for your decision with respect to the second version of our stage one registered report manuscript. We believe that the review process was once again timely and that the few remaining comments contributed to fine-tuning our manuscript.

To facilitate the review of the third version of our manuscript, comments are numbered and presented in table cells, below which are our responses. All revisions to the original version of our manuscript have been highlighted in green in the (.pdf) present document. All revisions to the original version of our manuscript have been highlighted using the "track changes" function in *Microsoft Word* in the (.docx) revised version of our manuscript.

Thank you for your involvement and, once again, for your positive appreciation!

Sincerely,

Loïs Fournier (corresponding author), on behalf of all the authors involved (Alexandre Heeren, Stéphanie Baggio, Luke Clark, Antonio Verdejo-García, José C. Perales, and Joël Billieux)

Loïs Fournier, M.Sc. (lois.fournier@unil.ch)
Institute of Psychology, University of Lausanne, Lausanne, Switzerland

Comment VMK-01

> The RQs certainly add clarity here. However, I suggest numbering them continuously or otherwise uniquely, which helps referring to each distinctly later (e.g., now there are three RQ1s and three RQ2s).

We acknowledge and agree with the present comment. In response, we have adopted a continuous numbering system for the research questions with respect to our three-phase development and evaluation protocol, thus ensuring that each research question is uniquely identified. We believe that this will also facilitate reference in the "Results" section of our stage two registered report manuscript.

In "2.1. Development phase I", the research questions now read as follows:

- "**RQ1.** What is the construct validity of the established 50-item version of the UPPS-P Impulsive Behavior Scale (UPPS-P-50)?"

- "**RQ2.** What is the internal consistency reliability of the established 50-item version of the UPPS-P Impulsive Behavior Scale (UPPS-P-50)?"

In "2.2. Development phase II", the research questions now read as follows:

- "**RQ3.** What is the content validity of the pre-established revised 20-item short version of the UPPS-P Impulsive Behavior Scale (UPPS-P-20-R)?"

- "**RQ4.** What is the construct validity of the pre-established revised 20-item short version of the UPPS-P Impulsive Behavior Scale (UPPS-P-20-R)?"

- "**RQ5.** What is the internal consistency reliability of the pre-established revised 20-item short version of the UPPS-P Impulsive Behavior Scale (UPPS-P-20-R)?"

In "2.3. Evaluation phase", the research questions now read as follows:

- "**RQ6.** What is the construct validity of the established revised 20-item short version of the UPPS-P Impulsive Behavior Scale (UPPS-P-20-R)?"

- "**RQ7.** What is the internal consistency reliability of the established revised 20-item short version of the UPPS-P Impulsive Behavior Scale (UPPS-P-20-R)?"

- "**RQ8.** What is the test-retest reliability of the established revised 20-item short version of the UPPS-P Impulsive Behavior Scale (UPPS-P-20-R)?"

- "**RQ9.** What is the criterion validity of the established revised 20-item short version of the UPPS-P Impulsive Behavior Scale (UPPS-P-20-R)?"

- "**RQ10.** What is the convergent validity of the established revised 20-item short version of the UPPS-P Impulsive Behavior Scale (UPPS-P-20-R)?"

Comment VMK-02

I agree with the reviewer that adding at least some threshold-like criteria at this point would further add credibility later. I understand no one wants to end up with a scale development paper that concludes the scale isn't valid and we don't want to create a prison that leads to such conclusions with a low bar; however, spelling out some boundaries of inference also makes your inference stronger later. You don't need criteria for each analysis but selecting at least some key RQs and setting them, e.g., gradual inference goals (X = adequate, Y = would benefit from improvement, Z = unacceptable, etc.) would be a huge epistemic jump. If it happens that some results aren't optimal, you can always continue by exploratorily testing with alternative data/models – it will be transparent for readers. In the end, we all want a useful scale with maximal awareness of its pros and cons. I'd love to see the field move even more away from valid/invalid semantics and be curious about different kinds and degree of utility (to be clear, I very much like it how you're already ahead of the main curve by assessing item as well as construct level nuance; it'd be great if you can extend such sensitivity to inference too).

We acknowledge and agree with the present comment, which echoes that of Ivan Ropovik (see "Comment IR-02"). In response, we have outlined fixed thresholds for inference criteria for research questions for which no inference criterion had been outlined in earlier versions of our manuscript. We believe that, in addition to providing explicit expectations regarding validity and reliability evidence in our stage one registered report manuscript, this will provide heuristics to inference that will be more straightforward to interpret in our stage two registered report manuscript.

Therefore, as per "Comment IR-02", in the corresponding sections of the revised version of our manuscript, we have added inference criteria that read as follows:

"Test-retest reliability evidence will be labeled "strong" if $\rho^2 \geq 0.500$, "adequate" if $\rho^2 \in [0.250, 0.500[$, "weak" if $\rho^2 \in [0.100, 0.250[$, and "absent" if $\rho^2 < 0.100$."

"Convergent validity evidence will be labeled "strong" if $\rho \geq 0.500$, "adequate" if $\rho \in [0.250, 0.500[$, "weak" if $\rho \in [0.100, 0.250[$, and "absent" if $\rho < 0.100$."

"Criterion validity evidence will be labeled "strong" if $\rho \geq 0.500$, "adequate" if $\rho \in [0.250, 0.500[$, "weak" if $\rho \in [0.100, 0.250[$, and "absent" if $\rho < 0.100$."

From: Loïs Fournier (corresponding author), on behalf of all the authors involved (Alexandre Heeren, Stéphanie Baggio, Luke Clark, Antonio Verdejo-García, José C. Perales, and Joël Billieux)

To: Ivan Ropovik (reviewer)

Comment IR-00

> After reading the revision of the Stage 1 RR, I see that the authors managed to further tighten the (already well-planned) design. Maybe a couple of outstanding issues.
>
> Thank you, best wishes,
>
> Ivan Ropovik

Dear Ivan Ropovik,

Thank you for reviewing the second version of our stage one registered report manuscript. We believe that the few remaining comments contributed to fine-tuning our manuscript.

To facilitate the review of the third version of our manuscript, comments are numbered and presented in table cells, below which are our responses. All revisions to the original version of our manuscript have been highlighted in green in the (.pdf) present document. All revisions to the original version of our manuscript have been highlighted using the “track changes” function in *Microsoft Word* in the (.docx) revised version of our manuscript.

Sincerely,

Loïs Fournier (corresponding author), on behalf of all the authors involved (Alexandre Heeren, Stéphanie Baggio, Luke Clark, Antonio Verdejo-García, José C. Perales, and Joël Billieux)

Loïs Fournier, M.Sc. (lois.fournier@unil.ch)
Institute of Psychology, University of Lausanne, Lausanne, Switzerland

Comment IR-01

> Yes, FIML – being a ML method – cannot be implemented when you use WLSMV and treating the indicators as ordinal. So, there is a tradeoff. Either you choose to use all participant data or use a superior method to model those. I'd do both, one as the default analysis and one as a sensitivity analysis. The question is what to go for as a default. I recommend building in a contingency for that. If the dropout due to listwise deletion (as done by *lavaan*) is above certain threshold, you could favor doing FIML, if it were below, you would go for WLSMV + ordinal model. The thing is that even with harmless-looking absolute rate of missingness (say 5%), you can have relatively substantial double digit listwise dropout, depending on the distribution of the missingness. Then, for the research questions where the substantive interpretation would differ for these two (arbitrarily chosen) pathways, I think it would be reasonable to remain in doubt and treat the results as inconclusive.
>
> Regarding the following sentence: "With respect to data collection, as we will require that participants provide answers to all statements implemented in the full online survey and as we will exclude data from participants who will have failed to complete the full online survey, no missing data will arise." (p.9) … and later in the manuscript, the same "no missing data will arise". It is even hard to tell how wrong this statement is. Missing data is a counterfactual phenomenon – it represents the data that would have been observed under different conditions and which inherently cannot be directly observed or retrieved but can only be estimated or inferred based on the available data and assumptions about the missingness mechanism. Some participants will be missing because of the forced responding, other data will be missing because listwise deletion will exclude a participant if they have just a single missing data point. So, saying that forcing + listwise deletion = no missing data will arise is so much wrong. Getting rid of participants with nonzero missingness is not equal to having no missingness. Listwise deletion is relatively fine when the data are MCAR, but not when data are MAR or even NMAR (in that case imputation is still better than listwise deletion).

With respect to the first part of the present comment, we have carefully considered the tradeoff between model estimation options and missing data handling options in our data analysis plan. The tradeoff can be summarized as follows:

Option 1.

- *Model estimation*. Adopting a continuous approach (<u>not recommended</u>)
- *Missing data handling*. Adopting a full information maximum likelihood approach (<u>recommended</u>)

Option 2.

- *Model estimation*. Adopting an ordered categorical approach (<u>recommended</u>)
- *Missing data handling*. Adopting a listwise deletion approach (<u>not recommended</u>)

Regarding model estimation, should a continuous approach or an ordered categorical approach be adopted? Simulation studies strongly support that treating ordered categorical observed variables as continuous observed variables is problematic in structural equation modeling (Kline, 2023). For example, Rhemtulla et al. (2012) reported that treating observed variables with fewer than five response categories as continuous observed variables generated biased estimates for model-implied parameter values, ultimately resulting in erroneous model fit assessments, as exact, approximate, and local fit based on continuous data assumptions may not accurately reflect the model's quality of adjustment to the data with ordered categorical data (e.g., inflated approximate fit indices). Consistent with Rhemtulla et al. (2012), Li (2016) reported that treating observed variables as continuous observed variables generated biased estimates for model-implied parameter values. In the context of our three-phase development and evaluation protocol, all (but one) psychometric instruments with respect to which we will collect participant data include items that are scored on a four-point Likert-type scale (the items of the *13-item Brief Self-Control Scale* (BSCS-13; Tangney et al., 2004) are scored on a five-point Likert-type scale). Therefore, considering the number of response categories being fewer than five, a continuous approach to model estimation must not be adopted.

However, whereas missing data handling options are numerous for continuous observed variables, such options for ordered categorical observed variables across confirmatory factor analyses, network analyses, and correlation analyses are limited to listwise deletion and pairwise deletion.

All in all, with respect to the first part of the present comment, we decided that we must prioritize model estimation over missing data handling, therefore adopting an ordered categorical approach to model estimation and a listwise deletion approach to missing data handling.

With respect to the second part of the present comment, we acknowledge and agree with it, and thank you for pointing out our misstatement. Indeed, reporting that "With respect to data collection, as we will require that participants provide answers to all statements implemented in the full online survey and as we will exclude data from participants who will have failed to complete the full online survey, no missing data will arise." is incorrect and was omitted in the revised version of our manuscript. Missingness information will be reported in the "Results" section of the manuscript through (1) percentage frequency values of participant-wise missingness and (2) percentage frequency values of data-wise missingness.

Taken together, in response to the present comment, the corresponding sections of the revised version of our manuscript read as follows:

"With respect to data collection, we will require that participants provide answers to all statements implemented in the full online survey and adopt a listwise deletion approach to missing data handling by excluding data from participants who will have failed to complete the full online survey. Missingness information will be reported in the "Results" section of the manuscript through (1) percentage frequency values of participant-wise missingness and (2) percentage frequency values of data-wise missingness.".

Kline, R. B. (2023). *Principles and practice of structural equation modeling*.

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical structural equation modeling estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*(2), 271–324. https://doi.org/10.1111/j.0022-3506.2004.00263.x

Comment IR-02

In my view, the treatment of validity assessment remains weak even in the revised protocol. The authors argued here that “accounting for issues related to statistical power, a “nomological network” approach cannot be adopted”. Well, it is completely alright to consider the internal latent structure of the validated measure as the primary aim and power the study to that aim. That way, the simulations for the sample size determination need not be extended or revised. You are doing a good job with respect to the internal structure aspect of construct validity but on its own, this cannot establish substantive meaning of the underlying latent variable. This is the role of the nomological network. Currently, you are just planning to do bivariate correlations of weighted sum-scores and for the criterion/convergent constructs only resort to repetitively argue that these “present differential associations with […]”. First, not sure what you mean by “presents differential associations”. More importantly, I reiterate my question about how the validity evidence will be assessed. Right now, there is one specific research question tied to criterion validity and one to convergent validity. But the criteria for proclaiming that the evidence speaks in favor of the given validity type are absent. Ideally, the authors would made explicit the theoretical expectations of construct interrelationships within a nomological network and test it empirically. I understand authors’ concerns regarding the computational tractability of a model that would involve a structural model of 11 latents and 64 indicators. In my view, this can be overcome by using a two-stage estimation, specifically the structural-after-measurement (SAM) approach. There are ways how to model the indicators as ordinal within SAM when using *lavaan*, but even if you treated them as continuous (the differences with respect to substantive interpretations are usually negligible), the validation procedure would still be much more powerful. In SAM, the measurement models are being estimated separately from the structural part. The latter would thus include only 11 variables which should be feasible even with your sample size. What you gain is robustness to error propagation due to local misspecification (e.g., in some of the measurement models of the other constructs), while having all the benefits of a full SEM. This approach will allow you to see how your underlying construct functions within a far more comprehensive theoretical framework compared to an independent bunch of bivariate correlations.

You could then also set some a priori thresholds considering the probabilistic nature of the interrelationships between the constructs, saying what you regard as good, acceptable theoretical fit, etc. (e.g., in terms of proportion of relationships within the nomological net that panned out as expected). Overall, I think the design would benefit from having at least some straightforward minimum-threshold heuristics for saying the validity evidence is at least adequate. Now that you have included explicit RQs, I recommend going through each of those and have some sort of idea about formal way of arriving at synthesizing judgment (as many of the RQs are inherently complex). I get that you cannot have clear-cut criteria for everything in a validation but having at least minimum criteria for supporting the use of the measure would be proper to have.

Before addressing the primary points raised in the present comment, we would like to clarify a terminology issue. Specifically, our use of the term “composite factor score” was imprecise, as we will use arithmetic mean scores to represent the scores of the psychometric instruments. Therefore, the term “composite factor score” was replaced by the term “arithmetic mean score” in the revised version of our manuscript.

With respect to the first part of the present comment, we understand that there are advanced techniques such as the structural-after-measurement (SAM) approach. However, we believe that correlation analyses are a conventional, parsimonious, and robust procedure to assess test-retest reliability, convergent validity, and criterion validity evidence. Furthermore, as detailed in the recent article of Sijtsma et al. (2024), the use of sum scores (which are monotonically equivalent to arithmetic mean scores) in psychometrics is supported by substantial evidence demonstrating their robustness and utility in clinical and research settings without requiring the additional complexity of latent variables in structural equation models. All in all, in this perspective, we decided to maintain correlation analyses performed using two-sided Spearman's ρ rank correlation tests on the psychometric instruments' arithmetic mean scores, as this procedure balances convention, methodological rigor, parsimony, and practical accessibility.

With respect to the second part of the present comment, we acknowledge and agree with it. Thus, we have outlined fixed thresholds for inference criteria for research questions for which no inference criterion had been outlined in earlier versions of our manuscript. We believe that, in addition to providing explicit expectations regarding validity and reliability evidence in our stage one registered report manuscript, this will provide heuristics to inference that will be more straightforward to interpret in our stage two registered report manuscript.

Therefore, in the corresponding sections of the revised version of our manuscript, we have added inference criteria that read as follows:

"Test-retest reliability evidence will be labeled "strong" if $\rho^2 \geq 0.500$, "adequate" if $\rho^2 \in [0.250, 0.500[$, "weak" if $\rho^2 \in [0.100, 0.250[$, and "absent" if $\rho^2 < 0.100$."

"Convergent validity evidence will be labeled "strong" if $\rho \geq 0.500$, "adequate" if $\rho \in [0.250, 0.500[$, "weak" if $\rho \in [0.100, 0.250[$, and "absent" if $\rho < 0.100$."

"Criterion validity evidence will be labeled "strong" if $\rho \geq 0.500$, "adequate" if $\rho \in [0.250, 0.500[$, "weak" if $\rho \in [0.100, 0.250[$, and "absent" if $\rho < 0.100$."

Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024). Recognize the value of the sum score, psychometrics' greatest accomplishment. *Psychometrika*, *89*(1), 84–117. https://doi.org/10.1007/s11336-024-09964-7

Comment IR-03

> Towards the end of the manuscript, you repetitively use the same phrases. Yes, we need to tighten things in an RR, but a RR is still a special case of literature and that is to be read by humans (for now, at least), so I would keep that in mind.

We have deliberately written the "Methods" and "Results" sections in a systematic way to make them easier to read.

Supplementary revisions

Supplementary revision 1

In the second version of our stage one registered report manuscript, we reported that dynamic threshold values will be estimated for three model-implied approximate fit indices using the *R* package *dynamic* version 1.1.0 or later (Wolf & McNeish, 2022): the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA) (Kline, 2023). However, using the latter *R* package, estimation of dynamic threshold values for model-implied approximate fit indices pertains to the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) (Kline, 2023; Wolf & McNeish, 2022). In this light, we replaced the Tucker-Lewis index (TLI) with the standardized root mean square residual (SRMR) in the corresponding sections of the revised version of our manuscript. Moreover, such supplementary revision aligns with the recommendations of Kline (2023).

Kline, R. B. (2023). *Principles and practice of structural equation modeling*.

Wolf, M. G., & McNeish, D. (2022). *dynamic: dynamic fit indices cutoffs for latent variable models* (1.1.0) [Computer software]. https://cran.r-project.org/package=dynamic

Supplementary revision 2

The labels "convergent validity" and "criterion validity" were inverted in the corresponding sections of the revised version of our manuscript to align with the labels reported in the *Standards for educational and psychological testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.