We may not be measuring physical closeness in interpersonal relationships as reliably as we think

*A recommendation by **Moin Syed** ⓘD based on peer reviews by **Jacek Buczny** ⓘD and **Maanasa Raghavan** of the STAGE 2 REPORT:*

Cite this recommendation as:

Attachment and interpersonal relationships are a major subject of research and clinical work in psychology. There are, accordingly, a proliferation of measurement instruments to tap into these broad constructs. The emphasis in these measures tends to be on the emotional dimensions of the relationships—how people feel about their partners and the support that they receive. However, that is not all there is to relationship quality. Increasing attention has been paid to the physical and physiological aspects of relationships, but there are few psychometrically sound measures available to assess these dimensions. In the current study, Dujols et al. (2024) assessed the psychometric properties of the Social Thermoregulation and Risk Avoidance Questionnaire (STRAQ-1), a measure of physical relationships that targets social thermoregulation, or how physical proximity is used to promote warmth and closeness. The project consists of a thorough assessment of the measure's reliability over time—that is, the degree to which the measure assesses the construct similarly across administrations, in a sample of 183 French university students. The authors assessed the longitudinal measurement invariance and test-restest reliability of the STRAQ-1. Longitudinal measurement invariance across two time points was only found for two of the four subscales. Similarly, test-retest reliability varied by subscale, ranging from poor to good. Taken together, the study suggests caution in using the STRAQ-1 scale as a reliable measure of physical relationships. The study highlights the need for continued assessment of the reliability of widely used measures, particularly reliability over time, and serves as a model for a rigorous analytic approach for doing so. The Stage 2 manuscript was evaluated over two rounds of in-depth review, the first round consisting of comments from two reviewers and the second round consisting of a close read

by the recommender. Based on detailed responses to the reviewers' comments, the recommender judged that the manuscript met the Stage 2 criteria and therefore awarded a positive recommendation. **URL to the preregistered Stage 1 protocol:** `https://osf.io/pmnk2`

Level of bias control achieved: Level 3. *At least some data/evidence that was used to the answer the research question was accessed by the authors prior to Stage 1 IPA (e.g. downloaded or otherwise received) but the authors certify that they had not observed ANY part of the data/evidence until after Stage 1 IPA.* **List of eligible PCI RR-friendly journals:**

- Collabra: Psychology
- F1000Research
- International Review of Social Psychology
- Peer Community Journal
- PeerJ
- Royal Society Open Science
- Social Psychological Bulletin
- Studia Psychologica
- Swiss Psychology Open

References:

1. Dujols, O., Klein, R. A., Lindenberg, S., Van Lissa, C. J., & IJzerman, H. (2024). Test-Retest Reliability of the STRAQ-1: A Registered Report [Stage 2]. Acceptance of Version 10 by Peer Community in Registered Reports. `https://osf.io/preprints/psyarxiv/392g6`

Reviews

Evaluation round #1

DOI or URL of the preprint: `https://osf.io/preprints/psyarxiv/392g6`
Version of the preprint: `https://osf.io/preprints/psyarxiv/392g6`

Authors' reply, 13 June 2024

Dear managing board of PCI Registered Reports,

Thank you for the editorial letter regarding our manuscript, "Test-Retest Reliability of the STRAQ-1: A Registered Report".

In this letter, we provide point-by-point responses to the issues identified by you and the two reviewers. To review the changes we made in the manuscript, we provide a manuscript with track changes on it and another

one without for readability. We appreciate the feedback provided by you and the reviewers and believe that it greatly improved the quality of our manuscript. We hope that the new version of the manuscript will now meet the high standards for Stage II acceptance at PCI Registered Reports.

Dr. Moin Syed informed us that, in case of Stage II acceptance, our manuscript is eligible for publication in the journal Advances in Methods and Practices in Psychological Science (AMPPS). We verified the word count limit in AMPPS which is 5000 words. We have currently 5798 words in our manuscript. If this word limit poses a problem for publication, we propose to remove the exploratory analysis section (919 words) from the manuscript but still report it as supplementary material on the OSF page of our project.

Sincerely, on behalf of all co-authors,
 Olivier Dujols

PCI RR Recommender: Dr. Moin Syed

 Point 1: Both reviewers felt that the conclusions drawn in the paper and abstract paint too rosy a picture of the psychometric properties of the instrument given the results (especially the measurement invariance findings). It is critical that the conclusions made are properly calibrated.

 Authors' Response: We changed the conclusion in the abstract. We replaced *"We discuss our findings in regard to the relatively long time between the repeated measures."* by *"Our study suggests that test-retest reliability was insufficient for psychological diagnosis, and that future studies should address the problem of low generalizability."* (lines 60-62 in the version without track changes).

 We also changed the conclusion in the conclusion section of the main text. We replaced *"Our results suggest a relative stability over time of the STRAQ-1 subscales and tend to support previous conceptualisation of the STRAQ-1 as a trait measure of individual differences in physical safety."* by *"Our study suggests that test-retest reliability was insufficient for psychological diagnosis, and that future studies should address the problem of low generalizability."* (lines 490-491 in the version without track changes).

 We additionally made some minor changes throughout the paper to better calibrate the interpretations, which are noted in our other responses.

 Point 2: Many of the results are attributed to low statistical power, a point on which I am confused. The power analysis included in your Stage 1 paper, which continues to appear in this Stage 2 manuscript, indicates sufficient power for both the invariance tests and ICC analyses. Accordingly, I was unclear how you concluded that you had insufficient power. Some of the language in the paper indicates that you may be relying on observed/post-hoc power, but as has been highlighted in the literature this should not be done as those values are data-dependent. Thus, the issue of power needs to be substantially clarified and modified throughout the paper.

 Authors' Response: We kept only the power analysis that was originally included in your Stage 1 paper: indicating a-priori sufficient power for both the invariance and ICC analyses. We understand your point about the issues with post-hoc power and removed all the post-hoc power analyses in the manuscript and the R scripts. We hope the power analysis is now clearer.

 We accordingly changed our conclusion and discussion related to these post-hoc power analyses, and concluded in regard to the a-priori power analyses. We removed *"Due to power issue"* from our conclusion section (line 486 in the version without track changes).

 Point 3: The results of the measurement invariance tests are presented much too quickly and without sufficient detail, relegating most of the crucial information to the table. The fact that one of the subscales did not even show configural invariance and another did not show metric invariance is not clearly stated or fully considered in the interpretations (see point 1 above). Moreover, the text indicates that you would report other fit indices in addition to CFI, which I think is a good idea, but then the table only includes them for the

configural model. Including them for all models would be helpful.

Authors' Response: We added more information in the text about our measurement invariance analysis. We added in the text the level of invariance reached by each scale.We rewrote the result paragraph about measurement invariance: *"Out of the four STRAQ-1 subscales, two reached longitudinal scalar invariance across two-time points. Table 2 provides a complete description of the fits of all the models. Based on the results of the longitudinal CFA models, we considered longitudinally invariant the subscales that reached scalar invariance. The Social Thermoregulation (Configural-Metric ΔCFI: +.014; Metric-Scalar ΔCFI < .001) and High-Temperature Sensitivity (Configural-Metric ΔCFI: +.012; Metric-Scalar ΔCFI < .001) subscales met our criteria to reach scalar invariance, and thus are considered longitudinally invariant across two-time points. On the contrary, the Risk Avoidance and Solitary Thermoregulation subscales were considered longitudinally non-invariant across two-time points. The Risk Avoidance subscale failed to reach metric invariance (Configural-Metric ΔCFI = - .027). The configural model of the Solitary Thermoregulation subscale had insufficient fit to the data ($χ2$ = 158.05, CFI = .899, RMSEA = .061, 90% CI RMSEA = [.043, .077], SRMR = .072). Based on these analyses, the Social Thermoregulation and High-Temperature Sensitivity constructs are thus respectively similar across two-time points and their latent scores can be meaningfully compared in our dataset. The Risk Avoidance and Solitary Thermoregulation constructs are thus respectively dissimilar across two-time points and their latent scores comparison may not be meaningful in our dataset."* (lines 297-312 in the version without track changes).

We additionally considered these results and added more interpretation in the discussion and conclusion (see your point 1). We rewrote a paragraph in our discussion: *"We concluded that two of the STRAQ-1 (Social Thermoregulation and High-Temperature Sensitivity) out of the four subscales were longitudinally invariant across two-time points in our sample. The current data, suggest that test-retest reliability was insufficient for psychological diagnosis, and that future studies should address the problem of low measurement invariance (see COTAN standards, Evers et al., 2015). The development of new scales including more culturally suitable items may resolve the low generalizability pointed out by our analyses. The Social Thermoregulation, Risk Avoidance and Eating Questionnaire – 2 (STRAEQ-2, Dujols et al., 2024) includes new scales developed at 53 sites in 32 countries. The STRAEQ-2 is currently in validation and could potentially resolve the measurement invariance issues pointed out by the current study. Future studies should test for longitudinal measurement invariance of the STRAEQ-2."* (lines 428-438 in the version without track changes).

Thank you for pointing out the discrepancy between our text and the table. We corrected the mistake and we now report all the fit indices in the table for all the models.

Point 4: Although you seem to do a nice job indicating what analyses were and were not preregistered (thank you!), I always ask authors to include the statement, "All reported analyses were preregistered unless specified otherwise," and then make sure that is true throughout the paper.

Authors' Response: We added the sentence *"All reported analyses were preregistered unless specified other-wise"*. At the end of the first paragraph of the result section (line 238). We verified that it is true for all reported analyses in the paper.

Point 5: I am a big fan of footnotes, but this paper has too many, with some key details in footnotes that should be in the paper. I encourage you to go through and integrate as many of them as possible into the text. Also, footnote 21, "Our discussion will include a detailed Constraints On Generality (Simons et al., 2017)" is a leftover from the Stage 1 and should be removed.

Authors' Response: We integrated most of the footnotes in the text, we went from 21 footnotes to now 3 footnotes. We removed footnote 21, thank you for noticing the mistake.

Point 6: This is admittedly stylistic, but the information about the Stage 1 acceptance that is at the beginning of the Discussion section is typically located in the Method.

Authors' Response: We moved the information about the Stage 1 acceptance to the beginning of the Method section (line 161 in the version without track changes).

Reviewer 1: Dr. Jacek Buczny

Point 7: My major concern is the main conclusion and the fact that it does not seem supported by the data.

The major conclusion: "Our results suggest a relative stability over time of the STRAQ-1 subscales and tend to support previous conceptualisation of the STRAQ-1 as a trait measure of individual differences in physical safety." and the data-driven statements: (1) "Due to power issues, we concluded that none of the STRAQ-1 subscales were longitudinally invariant across two-time points", and (2) "we found that test-retest reliability was overall moderate to good" (lines 427-433).

Given low power, low generalizability, the lack of good longitudinal measurement invariance, and good test-retest at best reliability, if I were you, I (1) would not conclude anything specific about (1) the instrument (by the way, reliability is a characteristic of measurement, not an instrument), and (2) would rather conclude that the data suggested that the test-retest was insufficient for psychological diagnosis (cf. COTAN standards, https://psynip.nl/wp-content/uploads/2022/05/COTAN-review-system-for-evaluating-test-quality.pdf), but (3) I would conclude that the findings are promising and new data must be collected to deal with the problem of low power and low generalizability in the first place.

Authors' Response: Thank you for your review and for sharing this reference with us. Accordingly to your comment, we did not conclude anything specific about the instrument but instead talked about the measurement made in our study. We removed the sentence *"Overall, our results are coherent with the previous findings in the literature."* in the first paragraph of the Discussion (line 580). We removed the sentence conclusion "*Our results suggest a relative stability over time of the STRAQ-1 subscales and tend to support previous conceptualisation of the STRAQ-1 as a trait measure of individual differences in physical safety."* We concluded that *"These findings are promising but our data suggest that test-retest reliability was insufficient for psychological diagnosis, and that future studies should address the problem of low generalizability of the constructs."* (lines 60-62 and 490-491 in the version without track changes).

Reviewer 2: Dr. Maanasa Raghavan

Point 8: This submission by Dujols and colleagues meets most of the Stage 2 criteria adequetely. One concern is the underpowered nature of the data. I commend the authors for presenting this result transparently in the Results and Discussion sections, but would suggest the authors consider paring back the overall conclusion of the study to reflect this limitation better.

Authors' Response: According to Moin Syed comment, we removed the post-hoc power analyses and kept the a-priori one that was on the Stage 1 manuscript. We adapted the overall conclusion accordingly (see points 1 to 3 and 7).

Point 9: N for participants with two data points switches between 183 and 184 throughout the text. Please check this and standardize, as needed.

Authors' Response: Thank you for pointing out the mistake. We corrected and inserted the correct number of participants throughout.

Point 10: Page 7, line 152: Should 2021-2020 perhaps be 2021-2022?

Authors' Response: Thank you again, we corrected the mistake, now saying *"2021-2022"*. (line 169 now in the version without track changes).

Download tracked changes file

Decision by Moin Syed ⬤, posted 10 May 2024, validated 11 May 2024

Stage 2 Decision Round #1: Minor Revision

May 10, 2024

Dear Authors,

Thank you for submitting your Stage 2 manuscript, "Test-Retest Reliability of the STRAQ-1: A Registered Report," to PCI RR.

First, please accept my apology for the extreme delay in sending this decision letter. I have had one review in hand for quite some time but was waiting for another. Once it became clear that the review would not be

coming, we worked to elicit a review from a member of the PCI RR board. Regardless of the reasons, the delay was simply too long and frankly unacceptable, so once again I apologize.

The good news is that the reviewers and I were all in agreement that your Stage 2 manuscript was well-prepared, and that it requires only a few revisions before I can issue a recommendation.

1. Both reviewers felt that the conclusions drawn in the paper and abstract paint too rosy a picture of the psychometric properties of the instrument given the results (especially the measurement invariance findings). It is critical that the conclusions made are properly calibrated.

2. Many of the results are attributed to low statistical power, a point on which I am confused. The power analysis included in your Stage 1 paper, which continues to appear in this Stage 2 manuscript, indicates sufficient power for both the invariance tests and ICC analyses. Accordingly, I was unclear how you concluded that you had insufficient power. Some of the language in the paper indicates that you may be relying on observed/post-hoc power, but as has been highlighted in the literature this should not be done as those values are data-dependent. Thus, the issue of power needs to be substantially clarified and modified throughout the paper.

3. The results of the measurement invariance tests are presented much too quickly and without sufficient detail, relegating most of the crucial information to the table. The fact that one of the subscales did not even show configural invariance and another did not show metric invariance is not clearly stated or fully considered in the interpretations (see point 1 above). Moreover, the text indicates that you would report other fit indices in addition to CFI, which I think is a good idea, but then the table only includes them for the configural model. Including them for all models would be helpful.

4. Although you seem to do a nice job indicating what analyses were and were not preregistered (thank you!), I always ask authors to include the statement, "All reported analyses were preregistered unless specified otherwise," and then make sure that is true throughout the paper.

5. I am a big fan of footnotes, but this paper has too many, with some key details in footnotes that should be in the paper. I encourage you to go through and integrate as many of them as possible into the text. Also, footnote 21, "Our discussion will include a detailed Constraints On Generality (Simons et al., 2017)" is a leftover from the Stage 1 and should be removed.

6. This is admittedly stylistic, but the information about the Stage 1 acceptance that is at the beginning of the Discussion section is typically located in the Method.

When submitting a revision, please provide a cover letter detailing how you have addressed the reviewers' points.

Thank you for submitting your work to PCI RR, and I look forward to receiving your revised manuscript.

Moin Syed

PCI RR Recommender

Reviewed by Jacek Buczny , 06 February 2024

Dear Authors,

Thank you for the detailed responses to my comments. Your analyses are excellent; congratulations!

My major concern is the main conclusion and the fact that it does not seem supported by the data. The major conclusion: "Our results suggest a relative stability over time of the STRAQ-1 subscales and tend to support previous conceptualisation of the STRAQ-1 as a trait measure of individual differences in physical safety." and the data-driven statements: (1) "Due to power issues, we concluded that none of the STRAQ-1 subscales were longitudinally invariant across two-time points", and (2) "we found that test-retest reliability was overall moderate to good" (lines 427-433).

Given low power, low generalizability, the lack of good longitudinal measurement invariance, and good test-retest at best reliability, if I were you, I (1) would not conclude anything specific about (1) the instrument (by the way, reliability is a characteristic of measurement, not an instrument), and (2) would rather conclude that the data suggested that the test-retest was insufficient for psychological diagnosis (cf. COTAN standards,

), but (3) I would conclude that the findings are promising and new data must be collected to deal with the problem of low power and low generalizability in the first place.

Good luck with your projects!

Jacek Buczny

Reviewed by Maanasa Raghavan, 09 May 2024

This submission by Dujols and colleagues meets most of the Stage 2 criteria adequetely. One concern is the underpowered nature of the data. I commend the authors for presenting this result transparently in the Results and Discussion sections, but would suggest the authors consider paring back the overall conclusion of the study to reflect this limitation better. Additionally, a couple of minor points:

- N for participants with two data points switches between 183 and 184 throughout the text. Please check this and standardize, as needed.

- Page 7, line 152: Should 2021-2020 perhaps be 2021-2022?

All the best!