**Test-Retest Reliability of the STRAQ-1: A Registered Report**

**Olivier Dujols1**

**Richard A. Klein2**

**Siegwart Lindenberg2,3**

**Caspar J. Van Lissa2**

**Hans IJzerman1,4**

1. Université Grenoble Alpes, France;

2. Tilburg University, The Netherlands;

3. Rijksuniversiteit Groningen, The Netherlands;

4. Institut Universitaire de France (IUF), France

***Author Note: We are submitting a Registered Report, but we wrote the manuscript in the past tense, even if the analyses have not been conducted yet.*** *The analyses we are yet to conduct are insert in brackets and are highlighted, to facilitate changes and review. We will insert and conclusions after results are finalized.*

***RRs involving existing data at PCI-RR:*** *Our project involves existing data. We estimate that we fall under the Level 3 concerning the dataset: “At least some data/evidence that will be used to answer the research question has been previously accessed by the authors (e.g., downloaded or otherwise received), but the authors certify that they have not yet observed ANY part of the data/evidence”. We (the first author) have already received the data, we certify that we have not observed any part of the evidence that could potentially be related to our research questions. We have only merged four datasets to compute a power analysis on the sample we will use for the analysis so the only variables from the dataset we consulted are: (1) the number of participants and (2) the participants’ anonymous codes.*

***Author Contributions:***

***Conceptualization:*** *Olivier Dujols, Richard Klein, Siegwart Lindenberg, Hans IJzerman*

***Data curation:*** *Olivier Dujols*

***Formal analysis:*** *Olivier Dujols and Hans IJzerman.*

***Funding acquisition:*** *Hans IJzerman and Olivier Dujols*

***Investigation:*** *Hans IJzerman and**Olivier Dujols*

***Methodology:*** *Olivier Dujols and Hans IJzerman*

***Project administration:*** *Hans IJzerman and Olivier Dujols*

***Resources:*** *Olivier Dujols and Hans IJzerman*

***Software:*** *Olivier Dujols*

***Supervision:*** *Hans IJzerman, Richard Klein, Siegwart Lindenberg*

***Validation:*** *Caspar van Lissa*

***Visualization:*** *Olivier Dujols*

***Writing - original draft:*** *Olivier Dujols and Hans IJzerman*

***Writing - review & editing:*** *Olivier Dujols, Richard Klein, Siegwart Lindenberg and Hans IJzerman*

**Abstract**

This Registered Report provides the first test of measurement invariance across time points and estimates of test-retest reliability for the Social Thermoregulation, Risk Avoidance Questionnaire (STRAQ-1, Vergara et al., 2019). The scale was developed and validated to understand the physiological drives underlying interpersonal bonding, measured by four constructs: the desire to socially regulate one’s temperature, the desire to solitary regulate one’s temperature, the sensitivity to higher temperatures, and the desire to avoid risk. Previous studies with large samples across 12 countries showed that the STRAQ-1 has a stable factorial structure, satisfying internal consistencies for the temperature subscales, and expected correlations in its nomological network. However, to date, this instrument has no estimates of test-retest reliability. Throughout four academic years (from 2018 to 2021), *N* = 184 French student participants took the STRAQ-1 at least two times. Out of the four STRAQ-1 subscales, X were longitudinally [non-invariant/invariant][[1]](#footnote-2) across two-time points. The constructs and latent scores were thus [dissimilar/similar][[2]](#footnote-3) and [incomparable/comparable][[3]](#footnote-4) across time. We then conducted test-retest reliability using Intra Class Correlation coefficient (ICC) for the *Social Thermoregulation*, *Solitary Thermoregulation*, *High-Temperature Sensitivity*, and *Risk Avoidance* subscales. ICCs estimates were respectively for agreement and consistency: XX, XX overall [excellent/good/moderate/poor][[4]](#footnote-5), XX, XX overall [excellent/good/moderate/poor], XX, XX overall [excellent/good/moderate/poor], and XX, XX overall [excellent/good/moderate/poor], respectively. We discuss our findings in regard to the relatively long time between the repeated measure (minimum one year).

***Keywords:*** *Test-Restest, Longitudinal Measurement Invariance, Attachment Theory, Social Thermoregulation, Registered Report*

**Study Design Table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question** | **Hypothesis** | **Sampling plan** | **Analysis Plan** | **Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis** | **Interpretation given different outcomes** | **Theory that could be shown wrong by the outcomes** |
| 1. Are each of the four STRAQ-1 subscales similar and comparable across two time points?  | We predict that longitudinal (configural, metric, and scalar) invariance holds across two-time points for all but one (*Risk Avoidance*).  | In this project, we relied on secondary data that was provided to us. A pool of psychology students replied to the STRAQ-1 at the beginning of four academic years, in 2018, 2019, 2020, and 2021 (*N\_2018* = 505, *N\_2019* = 298, *N\_2020* = 236, *N\_2021* = 400). We merged the participant responses across the four academic years based on a pseudo-anonymized code. In total, *N* = 184 French students took the STRAQ-1 at least two times. | For each subscale, we will compute a Confirmatory Factor Analysis, including the construct at T1 and T2 in the model. | We would need 10 participants per item included in the model to conduct the desired analysis. Our largest model will include 16 items and require a sample size of *N* = 160 participants (Kline, 2016). Our number of participants will be sufficient to conduct the analysis (*N* = 184 participants took the STRAQ-1 at least two times). We also computed an a priori sensitivity power analysis for the four STRAQ-1 configural models (but not specific power for metric, scalar, and residual invariance). The result of this analysis was that 164 participants would be required for our largest model. | We will run the tests for each of the subscales of the STRAQ-1. If one fails before reaching scalar invariance, we will consider the scores of this specific subscale not to be comparable across time points.  | Our theory is that three of the four subscales (*Social Thermoregulation, Solitary Thermoregulation*, *High-Temperature Sensitivity*) are comparable across time. If our test fails for those three measures, the measurement is not comparable across time. For the fourth (*Risk Avoidance*), because prior studies have shown that the reliability of the subscale was low, we expect it not to be comparable over time. If it is stable, it will disconfirm our a priori expectations and our theory will be updated.Explanations for all results will be presented in the discussion. |
| 2. Are each of the four STRAQ-1 subscale scores correlated and stable across time? | Based on previous studies that only investigated the reliability of the STRAQ-1, we expect that the test-retest reliability of the STRAQ-1 will be between moderate (Intra Class Correlation - ICC - between 0.5 and 0.75) and good (ICC between 0.75 and 0.9). | Same as above. | To assess test-retest reliability, we computed and reported ICC(2,1), to evaluate absolute agreement between the two time points, and ICC(3,1), to evaluate consistency. Both of these ICCs are calculated through 2-way mixed models. The STRAQ-1 subscales’ test-retest reliability between the 2-time points was estimated with intraclass correlation coefficients (ICCs) using the psych package in R (Revelle,2018). (Koo & Li 2016). | Again based on the number of participants (*N* = 184) with STRAQ-1 measurements over to two-time points, we have 80% power to detect an ICC of 0.2 (Bujang & Baharum, 2017). We will thus have sufficient power to detect our expected ICC between 0.5 and 0.9).Because researchers have argued that detection of non-zero ICC scores may not be sufficient and meaningful (see for example, Parsons, Kruijt & Fox, 2019), we also conducted a power analysis to estimate the 95% CI width that our sample will provide as a function of different ICC. This power analysis suggested that based on our sample size N = 184, we could estimate any ICC above .30 with a 0.2 width of the 95%CI, and any ICC above .80 with a 0.1 width of the 95% CI. | The measure of physical safety (or some of its components) is stable across time points separated by one year minimum.  | We theorize that three of the four subscales are comparable and stable over time. Comparability is tested under Question 1. For the ICC, we expect people's response to be reasonably stable. If this is not the case, then either (1) the measurement is not valid, or (2) people's personality on these measures is more dynamic than we had expected. In either case, future studies will then be needed to better understand which it is. We will make recommendations in the discussion if test-retest reliability is insufficient. |

**Test-Retest Reliability of the STRAQ-1: A Registered Report**

In the psychological literature, how people engage in interpersonal relationships is often understood through the prism of attachment theory, which proposes that individuals seek relational closeness to feel secure (Bowlby, 1969). But, while the importance of the physical safety of human infants is recognized in infant care in hospitals (e.g., temperature regulation), much less attention has been devoted to its self-report measurement in adults. Indeed, adult attachment measures focus primarily on the self-reported feelings of emotional safety and leave aside the issue of physical safety (e.g., Brennan, Clark, & Shaver, 1998; Fraley et al., 2000).

A notable exception to this is the Social Thermoregulation and Risk Avoidance Questionnaire (STRAQ-1) developed and validated by Vergara et al. (2019). The STRAQ-1 measures physical safety and the physiological drives underlying interpersonal bonding through four constructs: the desire to socially regulate one’s temperature, the desire to solitary regulate one’s temperature, the sensitivity to higher temperatures, and the desire to avoid risk. Previous studies in a large sample across 12 countries showed that the STRAQ-1 has a stable factorial structure, acceptable internal consistencies for the temperature subscales, and expected correlations in its nomological network. However, to date, no assessment of the test-retest reliability, crucial for the scale psychometrics and future use (e.g., evaluation of the impact of an intervention), has been conducted. In this article, we first assess longitudinal measurement invariance of the STRAQ-1 across time points, followed by an analysis of the test-retest reliability.

**Attachment and its measurement**

Bowlby (1969) proposed that social relationships are essential and adaptive to a child's survival since they are not able to survive by themselves. He postulated that a motivational system - the behavioral attachment system - drives the child to seek protection and support from the adult through crying and clinging behaviors. This behavioral system binds the child to the caregiver(s) so that they become attachment figure(s). Based on the availability and reliability of the care, the child will construct a mental representation (a working model) of the ability of their attachment figure to provide security, that in return, will impact their behaviors and feelings of security (Bretherton & Munholland, 2008; De Wolff & van IJzendoorn, 1997). These attachment patterns - the extent to which the child is secure or insecure in its relationship with the attachment figure - have been found to vary between individuals, and to be relatively stable from infancy to adulthood - when the main attachment figure becomes the romantic partner (Dugan & Chris Fraley, 2022; Fraley, 2019; Fraley et al., 2021).

 To measure attachment and identify how children differ based on it, an initial three-style classification was derived from observations of children: avoidant attachment, secure attachment, and anxious attachment (Ainsworth, 1979). This classification was later expanded to include disorganized attachment (Main & Solomon, 1986, 1990). In adults, the most widely used and currently psychometrically most sound instrument for measuring adult attachment styles is the Experiences in Close Relationships Inventory (ECRI; Brennan, Clark, & Shaver, 1998), which has since been revised (Experiences in Close Relationships Revised, ECR-R; Fraley et al., 2000, measured, for instance through the level of agreement with the statement *“I am very comfortable being close to romantic partners.”*). However, neither the ECRI, the ECR-R, nor the measures that preceded these adult attachment scales considered physical safety, such as protection against the cold, which is one essential aspect of survival proposed by Bowlby (1969)[[5]](#footnote-6).

**Social-thermoregulation-based attachment**

The theory of social-thermoregulation-based attachmentwas based on observations of non-human animals that found that when the temperature decreases, both infants and adults tend to move closer to their conspecifics to save energy and increase survival fitness (for example, through huddling, see Gilbert et al., 2010). The importance of physical proximity has also been studied in humans, demonstrating a determining role of thermoregulation in newborns. For instance, Bystrova et al. (2007) found that the mother's temperature was related to that of their infant, and increased after the birth of the infant (and even more so with skin-to-skin contact and early breastfeeding). In adulthood, attachment moderates people’s responses to temperature: securely attached people think of their loved ones when they are cold (versus warm), whereas this effect flips for those who are insecurely attached (IJzerman et al., 2018; see also Rocha IJzerman, 2021).

But existing attachment measures often do not map onto the concept of social-thermoregulation-based attachment. To better measure inter-individual differences in the regulation of temperature and risk through social relationships, Vergara et al. (2019) developed the Social Thermoregulation and Risk Avoidance Questionnaire (STRAQ-1). Across 12 countries and 1,510 participants, they found that the STRAQ-1 had a four-factor structure: (1) *Social Thermoregulation* (5 items; ωt = .83; reflecting the desire to warm up physically with close others), (2) *Solitary Thermoregulation* (8 items; ωt = .77; reflecting a desire to regulate temperature alone), and (3) *High-Temperature Sensitivity* (7 items; ωt = .83; reflecting a preference for colder temperatures and a distaste for hotter temperatures, and (4) *Risk Avoidance* (3 items; ωt = .57; reflecting the tendency to avoid - social - exploration)[[6]](#footnote-7).

In several French samples, the internal consistencies of the subscales were similar to those of the original validation study (Sarda et al., 2021; Vidal et al., 2022; Wittman et al., 2022[[7]](#footnote-8)). Vergara et al. (2019) also investigated the nomological network of the STRAQ-1[[8]](#footnote-9). We provide the most relevant correlations in Table 1. Again, the correlations with attachment have been replicated (excluding the *Risk Avoidance* subscale) in a French sample, and showed a similar pattern, with the addition of a relationship to loneliness (Wittman et al., 2022). However, despite evidence of the STRAQ-1 factorial structure, sufficiently high internal consistencies (except for the risk subscale), and validity through the nomological network, to date, no test-retest reliability has been conducted.

Test-retest reliability is crucial for scale psychometrics and for its future use. Without acceptable test-retest reliability, it is possible to confound artifacts of the measurement with true pre- and post-intervention differences in the rating of the scale, or miss the true effects of an intervention. Thus, test-retest reliability is necessary for theory development and to use the scale for interventions (cf., IJzerman, et al., 2017). Therefore the main purpose of this article is to examine the test-retest reliability of the STRAQ-1. Before doing so, we will provide an assessment of longitudinal measurement invariance of the STRAQ-1 across time points as it is a prerequisite for test-retest analysis[[9]](#footnote-10). This research was conducted in line with the CO-RE Lab Lab Philosophy v6 (Goncharova et al., 2022).

**Table 1.**

*Correlation in the nomological network of STRAQ-1.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Attachment Anxiety | Attachment Avoidance | Health | Stress | Self Control | Network Size |
| Social Thermoregulation | *n.s.* | -.31 | .10 | *n.s.* | *n.s.* | .10 |
| Solitary Thermoregulation | .08 | *n.s.* | *n.s.* | .11 | *n.s.* | *n.s.* |
| High-Temperature Sensitivity | .10 | *n.s.* | -.11 | .15 | -.17 | -.10 |
| Risk Avoidance | .17 | *n.s.* | -.11 | .24 | *n.s.* | -.12 |

*Note.* In the table, the reported correlations are all significant, the interested reader can refer to the Supplementary Materials via our OSF page (<https://osf.io/86qdx>) for the complete nomological network of the scale investigated in the original development paper (Vergara et al., 2019).

**Method**

**Participants**

A pool of psychology students replied to the STRAQ-1 at the beginning of four academic years, in 2018, 2019, 2020, and 2021 (*N\_2018* = 505, *N\_2019* = 298, *N\_2020* = 236, *N\_2021* = 400) as part of a larger “test week”. We merged the participant responses across the four academic years based on their pseudo-anonymized code. In total, *N* = 184 French students took the STRAQ-1 at least two times (xx males, xx females, xx unreported, *Mage* = xx, *SDage* = xx; *Mheight* = xx, *SDheight* = xx), from which *N* = 27 participants took the STRAQ-1 at least three times (xx males, xx females, xx unreported, *Mage* = xx, *SDage* = xx; *Mheight* = xx, *SDheight* = xx), and *N* = 4 participants took the STRAQ-1 four times (xx males, xx females, xx unreported, *Mage* = xx, *SDage* = xx; *Mheight* = xx, *SDheight* = xx)[[10]](#footnote-11).

**R packages**

We used the following R packages to conduct the analysis: rio (Chan et al., 2021), janitor (Firke, 2021), tidyverse (Wickham et al., 2019), psych (Revelle, 2022), GPArotation (Coen et al., 2005), EFA.dimensions (O'Connor, 2022), lavaan (Rosseel, 2012), semPlot (Epskamp, 2022), semTools (Jorgensen, 2021), energy (Rizzo & Szekely, 2022), semPower (Moshagen, & Erdfelder, 2016), ICC.Sample.size (Zou, 2012).

**Power analysis**

As we relied on secondary data, we did not conduct an a priori power analysis, but instead we conducted a sensitivity power analysis. Based on the number of participants (N = 184) that answered the STRAQ-1 at least twice, we calculated projected power to detect desired effect size. There are two recommendations for sample size for longitudinal measurement invariance analyses: five (Dimitrov, 2014) versus ten (Kline, 2016). In the former, we would need 5 (participants) \* 8 (items) \* 2 (time points) = 80 participants. In the latter, we would need 10 (participants) \* 8 (items) \* 2 (time points) = 160 participants. We also computed power for a general configural longitudinal measurement invariance models (CFA) models. We set power to 80%, alpha to .05, the amount of misfit to correspond to an RMSEA of at least .05, and the degrees of freedom to 100. The result of this analysis was that 164 participants would be required. In either case, our sample size was slightly above the required sample for detecting longitudinal measurement invariance over two time points.

We relied on Intra-Class Correlation (ICC) estimates for the test-retest reliability. Given that we had 184 participants with at least two time points, we had 80% power to detect an ICC of 0.2 with a pre-specified value of alpha of 0.05 (Bujang & Baharum, 2017). This means that if a small test-retest reliability exists (ICC = 0.2) we would have an 80% chance to detect it. However, we expected our subscales to present test-retest reliability between moderate (ICC between 0.5 and 0.75) and good (ICC between 0.75 and 0.9). Because researchers have argued that detection of non-zero ICC scores may not be sufficient and meaningful (see for example, Parsons, Kruijt & Fox, 2019), we also conducted a power analysis to estimate the 95% CI width that our sample will provide as a function of different ICC values.[[11]](#footnote-12) This power analysis suggested that based on our sample size *N* = 184, we could estimate any ICC above .30 with a 0.2 width of the 95% CI, and any ICC above .80 with a 0.1 width of the 95%CI. Hence, we had sufficient power to detect our expected ICC.

**Measure**

Participants rated the four subscales of the questionnaire STRAQ-1 on a Likert type scale ranging from 1 = Strongly Disagree to 5 = Strongly Agree. The *Social Thermoregulation subscale* presented [acceptable/poor][[12]](#footnote-13) internal consistency: McDonald’s ωt = .XX (e.g., “I prefer to warm up with someone rather than with something”). The *Solitary Thermoregulation subscale* presented [acceptable/poor] internal consistency: McDonald’s ωt = .XX (e.g., “When it is cold, I more quickly turn up the heater than others”). The *High-Temperature Sensitivity subscale* presented [acceptable/poor] internal consistency: McDonald’s ωt = .XX (e.g., “I am sensitive to heat”). The *Risk Avoidance subscale* presented [acceptable/poor] internal consistency: McDonald’s ωt = .XX (e.g., “I try to maintain myself in familiar places”). For each subscale, we averaged their items into a mean score.

**Results**

**Confirmatory Analyses**

The main goal of the analysis was to examine the test-retest reliability of the STRAQ-1, but longitudinal measurement invariance across time points must be established before conducting test-retest reliability (Chen, 2008). We first assessed the longitudinal measurement invariance of each of the four STRAQ-1 subscales across two-time points. We then ran the test-retest analysis. The R scripts of the analysis are available on the project’s OSF page: <https://osf.io/mr8n3/>.

**Longitudinal measurement invariance.** The main goal of the analysis was to ensure that the nature of the construct had not changed substantially over time. In longitudinal studies, the nature or meaning of a construct may change over time, resulting in longitudinal measurement non-invariance (Chen, 2008). Confirmatory Factor Analysis (CFA) is a common method for evaluating the level of invariance across time points (Widaman et al., 2010; Drasgow & Kanfer, [1985](https://link.springer.com/article/10.1007/s11469-019-00099-w)). Our procedure to test for longitudinal measurement invariance was to compare progressively more constrained CFA models. These models test incremental levels of measurement invariance across our two-time points (T1-T2). The levels of longitudinal measurement invariance have different implications for the construct: (a) if the configural level holds, then the structure of the measure is similar between T1 and T2; (b) if the metric level hold, then the structure of the measure and the constructs are similar between T1 and T2; (c) if the scalar level hold then the structure of the measure and the constructs are similar and the mean differences between T1 and T2 can be compared. Longitudinal scalar invariance is thus the minimal level required for our planned ICC analysis that uses the means scores of T1 and T2 (Kline, 2016; Mackinnon et al., 2022).

To investigate whether the variables in our dataset followed a multivariate normal distribution, we used the function `mvnorm.etest` from the Energy package. The analysis showed that our data [does/ does not] follow a multivariate distribution (E = XX, p = .XXX). A priori, we had already decided to use the WLSMV estimator instead of ML or MLR as arguments in the cfa function in lavaan to compute our CFA model, irrespective of the outcome of the test for multivariate normality. The WLSMV is the preferred solution when (a) the data is ordinal,[[13]](#footnote-14) and  (b) if data is potentially not normally distributed, as it makes no distribution assumptions (see Flora and Curran, 2004; Kline, 2016; Li, 2016). Then, we reported the robust weighted least squares fits for each model. We also verified the absence of Heywood cases (factor loading > 1 or negative variances) and residual correlations above *r* = .10. We then tested configural invariance, freely estimating the parameters and thresholds for T1 and T2, to verify whether the same latent factor structures held across time points. Our criteria for configural invariance were: comparative fit index < .95, root mean square error of approximation < .06 (CI 90% upper bond < .10, and non-significant *p*-value), and standardized root mean square residual < .05 (Kline, 2016).

Following our configural invariance test, we tested metric invariance, constraining the factor loadings and thresholds to be the same between T1 and T2, to verify whether the latent constructs were similar across time points. Then we tested scalar invariance, constraining the items’ intercepts and thresholds to be the same between T1 and T2, to ensure that the latent score at T1 and T2 were comparable. Finally, we tested residual invariance, further constraining the residual variances to be the same between T1 and T2, to ensure strict invariance of the latent score between T1 and T2.[[14]](#footnote-15)

To identify which level of longitudinal measurement invariance holds for each model, we followed the recommendation of [Mackinnon](https://osf.io/qbaah/) et al. (2022). Mackinnon et al. (2022) provided several criteria to access model fit for measurement invariance, one of these is the delta CFI (of .01) which is also recommended by a simulation study (Cheung & Rensvold, 2002). We decided to rely only on a ΔCFI of -.01 or more to conclude that the model with the largest CFI should be chosen. This means that if the ΔCFI is inferior or equal to -.01 we will choose the more parsimonious model and conclude for the longitudinal invariance of the specific level (metric, or scalar, or residual). Before pre-registration, we made choices about which metrics and cut-offs we would base our conclusion and interpretation of the subscale’s performance. But we acknowledge a lack of clear norms in the field about which metric to choose for our planned analyses. So, in addition to our pre-registered metric and cut-offs, we reported the results of other fit metrics even though we did not plan to use them for inferences and did not preregister any cut-of-value for them. This process will allow other researchers, who would prefer other indicators or cut-offs than ours, to be able to evaluate our models according to their criteria.

Out of the four STRAQ-1 subscales, X reached longitudinal [residual/scalar/metric/configural][[15]](#footnote-16) 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 the [*Social Thermoregulation*, *Solitary Thermoregulation*, *High-Temperature Sensitivity*, and *Risk Avoidance*][[16]](#footnote-17) subscale to be invariant across two time points.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model name | Configural model | Configural-Metric Δ fits | Metric-Scalar Δ fits | Decision about invariance |
| Social Thermoregulation | χ2 = CFI =RMSEA =SRMR = | ΔCFI =  | ΔCFI = | [configural/metric/scalar invariance/ non invariance] |
| Solitary Thermoregulation | χ2 =CFI =RMSEA =SRMR = | ΔCFI =  | ΔCFI = | [configural/metric/scalar invariance/ non invariance] |
| High Temperature Sensitivity | χ2 =CFI =RMSEA =SRMR = | ΔCFI = | ΔCFI = | [configural/metric/scalar invariance/ non invariance] |
| Risk Avoidance | χ2 =CFI =RMSEA =SRMR = | ΔCFI = | ΔCFI = | [configural/metric/scalar invariance/ non invariance] |

**Test-retest reliability.** The main goal of the analysis was to investigate the test-retest reliability of the four STRAQ-1 subscales (*Social Thermoregulation, Solitary Thermoregulation, High-Temperature Sensitivity,* and *Risk Avoidance*), using Intraclass Correlation Coefficient (ICC) analysis. The ICC analysis compares the variation across different ratings of the same individuals to the variation across all ratings and all individuals. An ICC close to 1 indicates that the scores from the same individual are highly similar. An ICC close to zero shows that the scores from the same individual are not similar. Koo & Li (2016) defined standards for the ICC with reliability being poor at ICC < 0.5; moderate at 0.5 < ICC > 0.75; good at 0.75 < ICC > 0.9; and excellent at ICC > 0.9. These are the cut-off values that we used for labeling our results. If the 95% confidence interval of an ICC estimate was in between two labels, we used both (for example, if the 95% CI interval would have been [0.83-0.94], the level of reliability would have been regarded as “good’ to “excellent”; see Koo & Li, 2016).[[17]](#footnote-19)

We computed and report ICC(2,1), to evaluate absolute agreement between the two time points, and ICC(3,1), to evaluate consistency. Both of these ICCs are calculated through two-way mixed-effect models. ICC(2,1) accounts for systematic and random error by specifying the time of measurement as a random effect in the model. ICC(3,1) only accounts for random error because the time of measurement is not specified as a random effect in the model (Koo & Li, 2016). The STRAQ-1 subscales’ test-retest reliability between the 2-time points was estimated with intraclass correlation coefficients (ICCs) using the psych package in R (Revelle, 2018). The analysis code is available on the OSF: <https://osf.io/mr8n3/>.

For the Sensitivity subscale, the estimated agreement was .XX, 95% confidence interval (CI) = [.XX, .XX], and the estimated consistency was .XX, 95% CI = [.XX, .XX]. For the Social Thermoregulation subscale, the estimated agreement was .XX, 95% confidence interval (CI) = [.XX, .XX], and the estimated consistency was .XX, 95% CI = [.XX, .XX]. For the Solitary Thermoregulation subscale, the estimated agreement was .XX, 95% confidence interval (CI) = [.XX, .XX], and the estimated consistency was .XX, 95% CI = [.XX, .XX]. Finally, for the Risk Avoidance subscale, the estimated agreement was .XX, 95% confidence interval (CI) = [.XX, .XX], and the estimated consistency was .XX, 95% CI = [.XX, .XX]. Out of X subscales we found X subscales to provide overall [excellent/good/moderate/poor][[18]](#footnote-20) test-retest reliability over two-time points.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

**Exploratory Analysis**.

We computed extra analyses that are labeled as exploratory either because of the relative degree of flexibility they introduce in the analysis (partial invariance), or because we did not have enough power to be sure of the effects (test-retest on more than two time points), or because we did not have a priori hypotheses but wanted to check the robustness of our confirmatory analyses (effect of the “academic year”).

**[Exploratory Partial Longitudinal Measurement Invariance.** We explored the partial longitudinal invariance of the scales that did not reach at least scalar longitudinal invariance in the confirmatory analysis. We found the subscales: X, X, and X, to reach respectively partial longitudinal [residual/scalar/metric/configural] invariance across two-time points.]

**Exploratory Intra Class Correlation.**

The next ICC analyses were exploratory because we did not have the power to test for longitudinal measurement invariance for three and four-time points. We computed ICCs estimates including only the 27 participants with three-time points. We computed the ICCs for *Social Thermoregulation*, *Solitary Thermoregulation*, *High-Temperature Sensitivity*, and *Risk Avoidance*. The ICCs were respectively XX, XX, XX, XX for agreement (ICC 2,1), and XX, XX, XX, XX for consistency (ICC 3,1)[[19]](#footnote-21), meaning that X subscales presented [excellent/good/moderate/poor][[20]](#footnote-22) [and X subscales presented excellent/good/moderate/poor] test-retest reliability across three-time points. We also computed ICCs in models including only the four participants that did the STRAQ-1 four-time. We computed the ICC for *Social Thermoregulation*, *Solitary Thermoregulation*, *High-Temperature Sensitivity*, and *Risk Avoidance*. the ICCs were respectively XX, XX, XX, XX for agreement (ICC 2,1), and XX, XX, XX, XX for consistency (ICC 3,1)[[21]](#footnote-24). These exploratory results indicated that X subscales presented [excellent/good/moderate/poor][[22]](#footnote-25) [and X subscales presented excellent/good/moderate/poor] test-retest reliability across more than two-time points.

**Exploratory effect of the academic year (over two-time points).** As a robustness analysis, we further computed the ICC estimates in a model including the “academic year” as a predictor of the STRAQ-1 scores over two-time points. Our rationale for this analysis was to investigate whether if the “academic year” could determine differences in STRAQ-1 scores (e.g., because of the onset of the COVID-19 pandemic or temperature changes over the years[[23]](#footnote-26)). These analyses were exploratory because we did not have a priori hypotheses about the effect of the academic year. Also, in case of an effect of the academic year, we would not have been able to say anything about the cause of the effect, and we would only have been able to speculate about why this effect occurred.

We used a linear mixed model to compute both ICCs estimates (ICC 2,1 and ICC 3,1) from four linear mixed models in which the academic year was a predictor of each of the STRAQ-1 scores. We computed the ICCs estimates for *Social Thermoregulation*, *Solitary Thermoregulation*, *High-Temperature Sensitivity*, and *Risk Avoidance[[24]](#footnote-27)*. We [found/did not find] a significant effect of the academic year on the [xxx subscales and xxx subscales] of the STRAQ-1, *b* = XX, *t(XX)* = XX, *p* = .XXX, *d* = XX. The ICCs were respectively XX, XX, XX, XX for agreement (ICC 2,1) and XX, XX, XX, XX for consistency (ICC 3,1)[[25]](#footnote-28). The results indicated that [minor/large] changes in the ICCs were induced from the models without the introduction of the academic year variable in the linear mixed models.

**Discussion[[26]](#footnote-29)**

**Conclusion**

**References**

Ainsworth, M. S. (1979). Infant–mother attachment. *American Psychologist, 34*(10), 932–937. DOI: [https://doi.org/10.1037/0003-066X.34.10.932](https://psycnet.apa.org/doi/10.1037/0003-066X.34.10.932)

Bowlby, J. (1969, 1982). Attachment and Loss, Vol. 1: Attachment. *Attachment and loss.* New York: Basic Books.

Brennan, K. A., Clark, C. L., & Shaver, P. R. (1998). Self-report measurement of adult romantic attachment: An integrative overview. In J. A. Simpson & W. S. Rholes (Eds.), *Attachment theory and close relationships* (pp. 46-76). New York: Guilford Press.

Bretherton, I., & Munholland, K. A. (2008). Internal working models in attachment relationships: Elaborating a central construct in attachment theory. In J. Cassidy & P. R. Shaver (Eds.), *Handbook of attachment: Theory, research, and clinical applications* (pp. 102–127). New York: Guilford Press.

Bujang, M. A., & Baharum, N. (2017). A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review. *Archives of Orofacial Science*, *12*(1).

Bystrova, K., Matthiesen, A. S., Vorontsov, I., Widström, A. M., Ransjö‐Arvidson, A. B., & Uvnäs‐Moberg, K. (2007). Maternal axillar and breast temperature after giving birth: effects of delivery ward practices and relation to infant temperature. *Birth*, *34*(4), 291-300. DOI: [https://doi.org/](https://psycnet.apa.org/doi/10.1037/0003-066X.34.10.932)[10.1111/j.1523-536X.2007.00187.x](https://doi.org/10.1111/j.1523-536x.2007.00187.x)

Chan, C., Chan, G. CH., Leeper, T. J., & Becker, J. (2021). Rio: A swiss-army knife for data file I/O. R package version 0.5.29.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology.* 95(1), 1005–1018. DOI: https://doi.org/[10.1037/a0013193](https://pubmed.ncbi.nlm.nih.gov/18954190/)

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. DOI: <https://doi.org/10.1207/S15328007SEM0902_5>

Coen A. B., & Robert I. J. (2005) Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement 65*(1), 676-696. DOI: https://doi.org/10.1177/0013164404272507

De Wolff, M.S. and van IJzendoorn, M.H. (1997), Sensitivity and Attachment: A Meta-Analysis on Parental Antecedents of Infant Attachment. *Child Development, 68*(1), 571-591. DOI: <https://doi.org/10.1111/j.1467-8624.1997.tb04218.x>

Dimitrov, D. M. (2014). *Statistical methods for validation of assessment scale data in counseling and related fields.* New York: John Wiley & Sons.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*(4), 662–680. DOI: [https://doi.org/10.1037/0021-9010.70.4.662](https://psycnet.apa.org/doi/10.1037/0021-9010.70.4.662)

Dugan, K. A., & Fraley, R. C. (2022). The roles of parental and partner attachment working models in romantic relationships. *Journal of Social and Personal Relationships*, *39*(7), 2154–2180. DOI: <https://doi.org/10.1177/02654075221075254>.

Epskamp, S. (2022). SemPlot: Path diagrams and visual analysis of various SEM packages' output. R package version 1.1.6.

Firke, S. (2021). Janitor: Simple tools for examining and cleaning dirty data. R package version 2.1.0.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. DOI: [https://doi.org/10.1037/1082-989X.9.4.466](https://doi.org/10.1037/1082-989x.9.4.466)

Fraley, R. C. (2019). Attachment in adulthood: Recent developments, emerging debates, and future directions. *Annual Review of Psychology*, *70*(1), 401-422. DOI: <https://doi.org/10.1146/annurev-psych-010418-102813>

Fraley, R. C., Dugan, K. A., Thompson, R. A., Simpson, J. A., & Berlin, L. J. (2021). The consistency of attachment security across time and relationships. In Thompson, R. A., Simpson, J. A., & Berlin, L. J. (Eds.). (2021). *Attachment: The fundamental questions*.New York: Guilford Press.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, *78*(2), 350-365. DOI: [https://doi.org/10.1037/0022-3514.78.2.350](https://psycnet.apa.org/doi/10.1037/0022-3514.78.2.350).

Gilbert, C., McCafferty, D., Le Maho, Y., Martrette, J. M., Giroud, S., Blanc, S., & Ancel, A. (2010). One for all and all for one: the energetic benefits of huddling in endotherms. *Biological Reviews*, *85*(3), 545–569. DOI: <https://doi.org/10.1111/j.1469-185X.2009.00115.x>

Goncharova, M., Silan, M. A., Dujols, O., Stoianova, T., Sparacio, A., Adetula, A., & IJzerman, H. (2019). The CO-RE Lab Lab Philosophy. *Available at* [*https://psyarxiv.com*](https://psyarxiv.com/6jmhe/download?format=pdf)*.*

IJzerman, H., Heine, E. C., Nagel, S. K., & Pronk, T. M. (2017). Modernizing relationship therapy through Social Thermoregulation Theory: Evidence, hypotheses, and explorations. *Frontiers in Psychology*, *8*(1), 635. DOI: https://doi.org/[10.3389/fpsyg.2017.00635](https://doi.org/10.3389/fpsyg.2017.00635)

IJzerman, H., & Neyroud, L., & Courset, R., & Schrama, M., & Post, J., & Pronk, T. (2018). Socially thermoregulated thinking: How past experiences matter in thinking about our loved ones. *Journal of Experimental Social Psychology. 79*(1). 349-355. DOI: https://doi.org/10.1016/j.jesp.2018.08.008.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). SemTools: Useful tools for structural equation modeling. R package version 0.5-5.

Kline, R. B. (2016). *Principles and practice of structural equation modeling (Fourth edition)*. New York: Guilford Press.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. DOI: <https://doi.org/10.1016/j.jcm.2016.02.012>

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(1), 936-949.DOI: <https://doi.org/10.3758/s13428-015-0619-7>

Mackinnon, S., Curtis, R., & O'Connor, R. (2022). A tutorial in longitudinal measurement invariance and cross-lagged panel models using lavaan. *Meta-Psychology, 6*(1). DOI: https://doi.org/10.31234/osf.io/tkzrb

Main, M., & Solomon, J. (1986). Discovery of an insecure-disorganized/disoriented attachment pattern. In T. B. Brazelton & M. Yogman (Eds.), *Affective development in infancy* (pp. 95-124). Westport: Ablex Publishing.

Main, M., & Solomon, J. (1990). Procedures for identifying infants as disorganized/disoriented during the Ainsworth Strange Situation. In *Attachment in the preschool years: Theory, research, and intervention* (pp. 121-160). Chicago: University of Chicago Press.

Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing structural equation models. *Structural Equation Modeling, 23*(1), 54-60. DOI: 10.1080/10705511.2014.950896

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

O'Connor B. P. (2022). EFA.dimensions: Exploratory Factor Analysis functions for assessing dimensionality. R package version 0.1.7.4.

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378-395. DOI: https://doi.org/10.1177/2515245919879695

Revelle, W. (2022) Psych: Procedures for personality and psychological research. Northwestern University, Evanston, Illinois, USA. R package version 2.2.9.

Rizzo, M., & Szekely, G. (2022). Energy: E-statistics, multivariate inference via the energy of data. R package version 1.7-10.

Rocha IJzerman, H. (2021). *Heartwarming: How our inner thermostat made us human*. New York: WW Norton & Company.

Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., ... & Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. *Journal of Abnormal Psychology*, *125*(6), 840. DOI: https://doi.org/10.1037/abn0000184

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(1), 1-36. DOI: https://doi.org/[10.18637/jss.v048.i02](https://doi.org/10.18637/jss.v048.i02)

Sarda, E., Klein, R. A., Dujols, O., & IJzerman, H. (2021). Validation of the ISP131001 sensor for mobile peripheral body temperature measurement. *International Review of Social Psychology*, *34*(1), 12. DOI : <https://doi.org/10.5334/irsp.409>

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. DOI: <https://doi.org/10.1177/1745691617708630>

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in Psychology*, *6*(1), 1064. DOI:  <https://doi.org/10.3389/fpsyg.2015.01064>

Vergara, R.C., Hernández, C., Jaume-Guazzini, F., Lindenberg, S., Klein, R.A., & IJzerman, H., 2019. Development and Validation of the Social Thermoregulation and Risk Avoidance Questionnaire (STRAQ-1). *International Review of Social Psychology*, *32*(1), 18. DOI: <https://doi.org/10.5334/irsp.222>

Vidal, N., Costello, J., Ribotta, B., Gurgand, L., & IJzerman, H. (2022). Assessing the reliability of an infrared thermography protocol to assess cold-induced Brown Adipose Tissue activation in young adults. *Preprint available at* [*https://psyarxiv.com/rkde9/download?format=pdf*](https://psyarxiv.com/rkde9/download?format=pdf)

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*(4), 319-350. DOI: https://doi.org/10.1016/j.jrp.2004.03.001

Wickham, H., Averick, M.,  Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Lin Pedersen, T., Miller, E., Milton Bache, S., Müller, K., Ooms, J., Robinson, D., Paige Seidel, D., Spinu, V., … & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. DOI: https://doi.org/10.21105/joss.01686

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10-18. DOI: <https://doi.org/10.1111/j.1750-8606.2009.00110.x>

Wittmann, A., Braud, M., Dujols, O., Forscher, P., & IJzerman, H. (2022). Individual differences in adapting to temperature in French students are only related to attachment avoidance and loneliness. *Royal Society Open Science*, *9*(5), 201068. DOI: <https://doi.org/10.1098/rsos.201068>

Zou, G. Y. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in Medicine, 31*(29), 3972-3981.

1. For a subscale to be labelled invariant, longitudinal scalar invariance had to hold and post hoc-power analysis had to show a power of 80 to detect scalar invariance. [↑](#footnote-ref-2)
2. Dissimilar: if metric invariance did not hold and/or if post-hoc power analysis did not show a power of 80 to detect metric invariance. Similar: if metric invariance held and if post-hoc power analysis showed a power of 80 to detect metric invariance. [↑](#footnote-ref-3)
3. Incomparable: if scalar invariance did not hold and/or if post-hoc power analysis did not show a power of 80 to detect scalar invariance. Comparable: if scalar invariance held and if post-hoc power analysis showed a power of 80 to detect scalar invariance. [↑](#footnote-ref-4)
4. To select the label for overall excellent/good/moderate/poor we will take the worst ICC between ICC(2,1) and ICC(3,1). Based on the cut off values defined by Koo & Li (2016) the label were: poor at ICC < 0.5; moderate at 0.5 < ICC > 0.75; good at 0.75 < ICC > 0.9; and excellent at ICC > 0.9. And if the 95% confidence interval of an ICC estimate was in between two labels, we used both (for example, if the 95% CI interval would have been [0.83-0.94], the level of reliability would have been regarded as “good’ to “excellent”). [↑](#footnote-ref-5)
5. See Vergara et al. (2019) for a more in-depth review of existing adult attachment measures and the motivation behind the development of the STRAQ-1. [↑](#footnote-ref-6)
6. The items of each subscale are presented in the Supplementary Materials (<https://osf.io/6px2u>). [↑](#footnote-ref-7)
7. The data in the Wittman et al. (2022) study are (partly) the same data that we are using in this project. The data have thus been previously observed by the main author Adrien Wittman and the person that did the code review (Mae Braud), but not in a way that would be related to the proposed analyses of this project. None of the analyses that we intend to conduct in the current have been run on the data. [↑](#footnote-ref-8)
8. All the reported correlations are significant, and the interested reader can refer to Vergara et al. (2019) for more details about all the correlations investigated in the original development paper. [↑](#footnote-ref-9)
9. In the Supplementary Materials (<https://osf.io/mr8n3/>) we also provide internal consistency (Alpha and Omega), per time point. We expected similar psychometrics in our current French samples compared to the original finding of Vergara et al. (2019). [↑](#footnote-ref-10)
10. Because we did not have specific hypotheses about the impact of specific academic years, we decided to label the first STRAQ-1 score that we had for a participant T1, with the second STRAQ-1 score T2 (and so forth). T1 and T2 thus do not reflect a specific academic year. The gap between T1 and T2 could vary between one to three years. For example, if a participant took the STRAQ-1 in 2019, 2020, and 2021, then this participant has three-time points: T1 then corresponds to the score of 2019, T2 to 2020, and T3 would be dropped from the pre-registered analysis (but would be included in the exploratory analysis). [↑](#footnote-ref-11)
11. The R code associated with this power analysis is available in the Supplementary Materials via our OSF page: <https://osf.io/mr8n3/>. [↑](#footnote-ref-12)
12. Our cut-off for the selection of the labels for internal consistency was: above or equal to .70 for acceptable, and under .70 for poor. This cut-off is often used in the literature and is based on Nunnally & Bernstein (1994) even if it was not intended as a gold standard for acceptable internal consistency. [↑](#footnote-ref-13)
13. Our measure is a 5-point Likert type scale, the label are (1) “Strongly disagree”, (2) “Disagree” (3)

“Neutral”, (4) “Agree”, (5) “Strongly agree”. But the numbers do not necessarily represent equal intervals or differences in magnitude between the ordered labels. Consequently, data obtained from a Likert scale are generally considered as ordinal, rather than continuous (where the intervals are equal between values). [↑](#footnote-ref-14)
14. Residual invariance has been described to be hard to reach for most psychological measurement instruments (Kline, 2016; van De Schoot et al., 2015). We thus considered longitudinally invariant the subscales that reached scalar invariance with sufficient power to detect the invariance. [↑](#footnote-ref-15)
15. For a label to be chosen the specific level of invariance had to hold and post hoc-power analysis had to show a power of 80 to detect the invariance. [↑](#footnote-ref-16)
16. We considered longitudinally invariant the subscales that reached scalar invariance with a power of 80 to detect the invariance. [↑](#footnote-ref-17)
17. We recognize that the discussion around cut-offs is contentious and that cut-offs are often arbitrarily chosen, which may make our values equally arbitrary (see e.g., Watson, 2004). The resulting labels (e.g., “good’) are considered as one of many means to assess the validity of a measure (Rodebaugh et al., 2016) and a first step towards defining a normative range of reliability estimates for a scale that will be applied across samples or contexts. [↑](#footnote-ref-19)
18. To select the label for overall excellent/good/moderate/poor we took the worst ICC between ICC(2,1) and ICC(3,1). Based on the cut off values defined by Koo & Li (2016) the label were: poor at ICC < 0.5; moderate at 0.5 < ICC > 0.75; good at 0.75 < ICC > 0.9; and excellent at ICC > 0.9. And if the 95% confidence interval of an ICC estimate was in between two labels, we used both (for example, if the 95% CI interval would have been [0.83-0.94], the level of reliability would have been regarded as “good’ to “excellent”). [↑](#footnote-ref-20)
19. For this test we had 90% power to detect an ICC of 0.4. [↑](#footnote-ref-21)
20. To select the label for overall excellent/good/moderate/poor we took the worst ICC between ICC(2,1) and ICC(3,1). Based on the cut off values defined by Koo & Li (2016) the label were: poor at ICC < 0.5; moderate at 0.5 < ICC > 0.75; good at 0.75 < ICC > 0.9; and excellent at ICC > 0.9. And if the 95% confidence interval of an ICC estimate was in between two labels, we used both (for example, if the 95% CI interval would have been [0.83-0.94], [↑](#footnote-ref-22)
21. For this test we had 80% power to detect an ICC of 0.8. [↑](#footnote-ref-24)
22. To select the label for overall excellent/good/moderate/poor we took the worst ICC between ICC(2,1) and ICC(3,1). Based on the cut off values defined by Koo & Li (2016) the label were: poor at ICC < 0.5; moderate at 0.5 < ICC > 0.75; good at 0.75 < ICC > 0.9; and excellent at ICC > 0.9. And if the 95% confidence interval of an ICC estimate was in between two labels, we used both (for example, if the 95% CI interval would have been [0.83-0.94], [↑](#footnote-ref-25)
23. Our mean and standard deviation of temperature in degrees Celsius (in degrees Fahrenheit in between parentheses) in Grenoble for the years included in the sample were for 2018 *Mtemp* = xx, *SDtemp* = xx (*Mtemp* = xx, *SDtemp* = xx); for 2019 *Mtemp* = xx, *SDtemp* = xx (*Mtemp* = xx, *SDtemp* = xx); for 2020 *Mtemp* = xx, *SDtemp* = xx (*Mtemp* = xx, *SDtemp* = xx); for 2021 *Mtemp* = xx, *SDtemp* = xx (*Mtemp* = xx, *SDtemp* = xx). In addition, the mean temperature in degrees Celsius (again, Fahrenheit in between parentheses) of the days we conducted the study were XX (XX) for 2018, XX (XX) for 2019, XX (XX) for 2020, XX (XX) for 2021. [↑](#footnote-ref-26)
24. As in the confirmatory analysis section, we excluded participants and consider them outliers only if their Cook's D or Lever presents “gaps” (value at least three times the Cook D or Lever of the previous value for the highest value) or Studentized residual absolute value was above four. [↑](#footnote-ref-27)
25. For this test. we had 90% power to detect an ICC of 0.4. [↑](#footnote-ref-28)
26. Our discussion will include a detailed Constraints On Generality (Simons et al., 2017). [↑](#footnote-ref-29)