Published in final edited form as:
J Pain. 2022 September ; 23(9): 1543–1555. doi:10.1016/j.jpain.2022.01.011.

Test-Retest Reliability of an Adaptive Thermal Pain Calibration Procedure in Healthy Volunteers

Carolyn Amir[#*], Margaret Rose-McCandlish[#*], Rachel Weger[#*], Troy C. Dildine[*,†], Dominik Mischkowski[‡], Elizabeth A. Necka[*,§], In-seon Lee[¶,‖], Tor D. Wager[**], Daniel S. Pine[††], Lauren Y. Atlas[*,††,‡‡]

[*] National Center for Complementary and Integrative Health, National Institutes of Health, Bethesda, Maryland

[†] Clinical Neuroscience Section, Karolinska Institutet, Solna, Sweden

[‡] Ohio University, Athens, Ohio

[§] National Institute on Aging, National Institutes of Health, Bethesda, Maryland

[¶] College of Korean Medicine, Kyung Hee University, Seoul, Republic of Korea

[‖] Acupuncture & Meridian Science Research Center, Kyung Hee University, Seoul, Republic of Korea

[**] Dartmouth College, Hanover, New Hampshire

[††] National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland

[‡‡] National Institute on Drug Abuse, National Institutes of Health, Baltimore, Maryland

[#] These authors contributed equally to this work.

Abstract

Quantitative sensory testing (QST) allows researchers to evaluate associations between noxious stimuli and acute pain in clinical populations and healthy participants. Despite its widespread use, our understanding of QST's reliability is limited, as reliability studies have used small samples and restricted time windows. We examined the reliability of pain ratings in response to noxious thermal stimulation in 171 healthy volunteers (n = 99 female, n = 72 male) who completed QST on multiple visits ranging from 1 day to 952 days between visits. On each visit, participants underwent an adaptive pain calibration in which they experienced 24 heat trials and rated pain intensity after stimulus offset on a 0 to 10 Visual Analog Scale. We used linear regression to determine pain threshold, pain tolerance, and the correlation between temperature and pain for each session and examined the reliability of these measures. Threshold and tolerance were moderately reliable (Intra-class correlation = .66 and .67, respectively; $P < .001$), whereas

Address reprint requests to Lauren Y. Atlas, PhD, National Institutes of Health (NCCIH, NIMH, NIDA), 10 Center Drive, Bethesda, MD 20892. lauren.atlas@nih.gov.

The authors report no conflicts of interest.

Supplementary data
Supplementary data related to this article can be found at https://doi.org/10.1016/j.jpain.2022.01.011.

temperature-pain correlations had low reliability (Intra-class correlation = .23). In addition, pain tolerance was significantly more reliable in female participants than male participants, and we observed similar trends for other pain sensitive measures. Our findings indicate that threshold and tolerance are largely consistent across visits, whereas sensitivity to changes in temperature vary over time and may be influenced by contextual factors.

Perspective: This article assesses the reliability of an adaptive thermal pain calibration procedure. We find that pain threshold and tolerance are moderately reliable whereas the correlation between pain rating and stimulus temperature has low reliability. Female participants were more reliable than male participants on all pain sensitivity measures.

Keywords

Test-retest reliability; quantitative sensory testing; thermal; healthy volunteers; intraclass correlations

Quantitative sensory testing (QST) is a valuable psychophysical tool for pain assessment in healthy volunteers[27] and in clinical populations.[7] QST complements bedside examinations by using standardized procedures to evaluate sensory thresholds and suprathreshold pain perception.[34] This permits comparisons with normative data and comparisons between individuals or across sessions within an individual. Understanding whether metrics are reliable is critical for evaluating QST's utility as a tool[58] for pain assessment and diagnosis. Reliability is particularly important when evaluating pain biomarkers or signatures,[55] as subjective pain is the gold standard against which any pain biomarker is compared.[19]

While QST reliability studies historically focus on warm and cool detection thresholds (for a review, see Moloney et al[52]), recent studies have evaluated the reliability of pain thresholds and suprathreshold pain perception in individuals with chronic pain[27,72] and healthy volunteers.[28,47] Most thermal pain reliability studies use the method of limits[25,35,36] to identify pain thresholds, ie the temperature at which a stimulus is labeled as painful, which corresponds to the activation of peripheral nociceptive C fibers.[40] Test-retest reliability estimates of heat pain threshold range from poor[77] to excellent[54,74]; however, these studies vary in quality as well as methodology,[52] with large variations in sample size, duration between sessions, and analytic approaches.

The method of limits is a clinically useful and fast procedure to detect warmth and pain thresholds, but it also has limitations, namely that it is dependent on reaction time, motivation, and attention.[34] The aim of the present study was to evaluate the test-retest reliability of QST measures staircase calibration (ASC[3,21,49]), in which participants provide pain ratings after heat offset and temperatures are selected based on an iterative regression; this task is similar to the method of levels.[76] Adaptive staircase calibrations are often used to select suprathreshold stimuli for use in subsequent experiments,[3,23,31,64,78] as they ensure participants can tolerate repeated stimulations at a given intensity. The ASC task allows researchers to simultaneously evaluate three independent measures of pain sensitivity: 1) pain threshold; 2) pain tolerance (the maximum pain an individual is willing to tolerate); and 3) the reliability of the association between temperature and pain. The ASC task was previously employed to select temperatures that were used to train the Neurologic Pain

Signature (NPS), a brain-based classifier that can predict whether a stimulus is painful or not.[73] The NPS has high reliability both within and across participants[33]; yet how reliable are the subjective pain measures that were used to train the NPS?

Our aim was to evaluate the test-retest reliability of the ASC task in healthy volunteers who completed the ASC task on multiple visits. We hypothesized that pain threshold, pain tolerance, and the strength of the temperature-pain association would be reliable across visits. We also conducted two exploratory analyses of factors that might impact the reliability of pain measures. First, we explored whether reliability is influenced by time between measurements, as previous studies of QST reliability have focused on either short-term reliability (eg 1−2 days between tests[28]) or long-term reliability (eg months between visits[47]). Our goal was to determine whether the duration between visits influences reliability, and we hypothesized that pain measures would be more stable when visits were closer in time. Second, we explored whether there are differences in reliability as a function of gender, in response to the need for greater emphasis on the measurement of sex as a biological variable.[1,17,18] Much pain research has excluded women and female animals from pain studies,[61–63] based on the assumption that hormonal fluctuations will be associated with greater variability in females. Yet this assumption has been shown to be erroneous in rodent studies in general[56] and in rodent models of pain.[51] To our knowledge, no previous study has evaluated gender differences in the test-retest reliability of pain sensitivity in humans, despite decades of work on sex differences in pain perception.[8] Understanding whether reliability differs as a function of gender is distinct from asking whether there are gender differences in tolerance, as reliability focuses on variability within individuals across time, and is a more direct test of the potential impact of hormonal fluctuations. Finally, to address limitations of previous work, we tested reliability of QST measures in a large sample (n = 171), compared multiple analytic approaches, and controlled for testing environment and visit number to test whether these experimental factors systematically affect pain sensitivity and thereby contribute to within-subject variability across sessions.

Methods

Participants

Participants were healthy volunteers screened from a community sample. Volunteers were recruited via clinicaltrials.gov, the National Institutes of Health (NIH) Office of Patient Recruitment, and flyers posted at the NIH Bethesda campus. Volunteers were ineligible if they had a history of psychiatric or neurological disorder, chronic pain (defined as pain lasting more than 6 months), substance abuse, or a major medical condition that could affect somatosensation. Participants were between the ages of 18 and 50, were fluent in English, and were not pregnant. During a nursing exam or medical exam prior to sensory testing, a nurse or clinician verified that participants had not taken any pain relievers within 5 half-lives.

Three hundred forty-two participants provided informed consent under an NIH IRB-approved protocol, through a general phenotyping and screening protocol (16-AT-0077; ClinicalTrials.gov Identifier: NCT02707029) and/or through a thematic protocol with multiple sub-studies (15-AT-0132; ClinicalTrials.gov Identifier: NCT02446262) that included an initial visit to evaluate eligibility. As specified in the protocols, participants could complete multiple visits and sub-studies and underwent an adaptive staircase calibration (ASC) with noxious heat on every visit to evaluate eligibility and identify temperatures for use in subsequent experimental procedures on that visit (see “Adaptive Staircase Calibration Procedure,” below). In the current paper, we focus on all participants who completed multiple heat calibrations administered on different days (n = 171). Data were collected between June 2015 and February 2020. Participants had an average age of 28.60 ± 7.83 years; 99 identified as female (57.90%) and 72 identified as male (42.11%). Ninety-two identified themselves as White (53.80%), 42 as Black/African American (24.56%), 27 as Asian (15.79%), 1 as Native Hawaiian (.58%), and 4 as more than one race (2.34%). Fourteen participants identified as Hispanic/Latino (8.19%), of whom 8 identified as White, 1 identified as Black/African American, and 5 did not report racial identity or marked it as “Unknown”.

Stimuli and Apparatus

Heat stimulation was administered via a 16 × 16 mm Advanced Thermal Stimulator thermode and controlled with a Pathway Pain and Sensory Evaluation System (Medoc Advanced Medical Systems Ltd, Ramat Yishay, Israel). Data were collected using multiple Pathways systems and thermodes were regularly calibrated to ensure temperature accuracy. Noxious stimulation ranged from 36°C to 50°C and is described in more detail below. The thermode was kept at a constant temperature of 32°C between stimulations. Pain ratings were collected either verbally or through a mouse and computer using a visual analogue scale (VAS).

Procedures

General procedures have been covered in detail in previous publications.[21,49] In brief, for each visit, participants provided informed consent, completed questionnaires, and then underwent an adaptive staircase calibration (ASC) task either in an outpatient testing room or in a suite adjacent to the functional magnetic resonance imaging (fMRI) scanner. On participants’ first visits, they underwent a nursing exam in the outpatient clinic to confirm eligibility prior to any procedures and received a physical exam if they had not had one at the NIH within the prior year to further screen for eligibility. The ASC was used to determine pain sensitivity and eligibility for subsequent testing in pain modulation experiments. We describe eligibility criteria below. The same ASC procedure was repeated on all visits, which occurred either in the clinic or in the MRI facility. The present analysis focuses only on participants who completed the ASC task on more than one visit (n = 171). Subsets of these data were included in previous papers on autonomic responses to heat,[49] relationships with trait mindfulness[50] and confidence in subjective pain.[21]

Adaptive staircase calibration procedure.—The ASC procedure has been described in detail in previous work.[3–5,21,49,50] All participants completed the ASC task during an initial screening visit, in which they received 24 trials of heat on the left volar forearm (see Fig 1). In brief, we applied 3 rounds of noxious heat to 8 skin sites, with temperatures determined through an iterative regression procedure. All stimuli were 8s including 3s ramping between baseline and target temperature, with the exception of 12 Screening Visit participants whose stimuli were 10s in duration during their first visit (see Supplementary Table S1 for study details). After the offset of each heat stimulus, participants rated it on a 0 to 10 VAS, where 0 was described as no sensation, 1 as warmth but not pain, 2 as pain threshold (the beginning of painful sensation), 5 as moderate pain, 8 as pain tolerance (the most amount of pain that participants were willing to tolerate), and 10 as the most pain imaginable. The instructions given to participants are provided in Dildine et al.[21] The first three temperatures were set the same for all participants (41°C, 44°C, and 47°C), and subsequent stimulus temperatures were selected using an iterative linear regression to identify temperatures predicted to elicit ratings of 2, 5, and 8 for each subject. To account for warmth-insensitive skin fields in the forearm,[32] outlier trials (defined as ratings that exceeded 2.5 times the median absolute deviation[42]) were excluded from the linear regression and from analyses of reliability.

Trial order was the same across participants to ensure that each site received stimuli at each predicted pain intensity (ie threshold, moderate pain, and maximum tolerable pain) and to ensure that each intensity was equally likely to be followed by every other intensity level. We also rotated through the sites across the duration of the task to avoid sensitization or habituation. All sites received predicted temperatures, rounded to the nearest .5°C, unless that temperature or a lower temperature had already been rated as intolerable (ie > 8) at that skin site, since participants were informed we would not apply stimuli they had deemed to be intolerable, consistent with our IRB-approved protocol and the IASP's guidelines for pain research in humans (https://www.iasp-pain.org/resources/guidelines/ethical-guidelines-for-pain-research-in-humans/). In this case, the experimenter manually lowered the temperature, usually to .5°C or 1°C below the ASC-predicted temperature. The regression was based on the applied temperature rather than the predicted temperature.

For each individual, we used linear regression to determine the temperature corresponding to level 2 as a measure of pain threshold and level 8 as a measure of pain tolerance. We estimated the correlation between temperature and pain and used r^2 as a measure of goodness-of-fit both with and without outliers. We note that the three pain sensitivity measures derived from the ASC task are related but independent: The relationship between pain threshold and pain tolerance depends on the slope of the regression line, and r^2 depends on the residuals between observed pain ratings and those that would be predicted by the regression.

Fig 1B depicts the study flow. As dictated by our clinical protocol, participants were ineligible for experimental tasks and subsequent visits if their ratings were not ordinally consistent with temperature or they had an r^2 value of less than .4 (based on calculation without outliers; n = 16 ineligible; 4.7% of participants screened), if they had a pain threshold below 36°C (n = 11; 3.22% of participants), pain tolerance above 50°C (n = 43; 12.57% of participants), or if there was a difference of less than 4°C between threshold and tolerance (n = 36; 10.6% of participants). One hundred seventy-one participants were deemed eligible for our follow up studies and completed more than one visit; 171 participants completed a single visit either because they were deemed ineligible for one or more of the reasons listed above (n = 88; 25.73% of participants screened) or due to scheduling constraints, lack of interest, or other reasons. Analyses comparing ASC measures between participants who completed a single visit with those who completed more than one visit are reported in Supplementary Materials and descriptives are included in Table 1 for completeness.

We focused on eligible participants who completed more than one ASC task across multiple visits and evaluated the reliability of thermal pain threshold, tolerance, and goodness-of-fit as estimated during the ASC task using linear regression that excluded outlier trials, as described above. In Supplementary Materials, we also report results from linear regressions that included outlier trials, and nonlinear models that account for potential nonlinearities between temperature and pain,[66,67] as well as information on the full sample.

Experimental and contextual procedures.—

Participants completed the ASC task during an initial visit and then completed the task on each subsequent visit to establish eligibility and select temperatures for use during that visit's experimental session. All visits took place at the NIH Clinical Center: Initial visits were conducted in a behavioral testing room in an outpatient clinic and subsequent visits could take place either in the testing room or in a suite adjacent to the MRI scanner. The basic ASC procedure was consistent across studies and visits with a few variations depending on the specific study for which the participant was being screened. Supplementary Table S1 reports study details and descriptive statistics for each study based on the entire sample as well as the number of participants who completed more than one visit and were included in the current analyses. In particular, heat stimuli lasted for 8 seconds in all visits except for 12 participants who experienced 10-second stimuli on their first visit, and ratings were provided either verbally or through a computer program. In one study, participants (n = 23) received heat stimuli on the left calf instead of the left forearm, in which case 12 sites were tested rather than 8. In another study, participants (n = 71) completed the ASC for heat on the left volar forearm and also rated pleasant or unpleasant liquid taste stimuli with a similar procedure; order of stimuli was counterbalanced across participants. These participants also rated pain unpleasantness as well as intensity. We focus on pain intensity ratings alone. Finally, different sessions may have been conducted by the same experimenter or different experimenters, the number of experimenters may have varied across sessions, and experimenters varied in gender, ethnicity, and race. See Supplementary Materials for an analysis of variance components as a function of subsequent study type.

Analyses

Evaluating reliability, repeatability, and agreement.—

Our main goal was to evaluate the test-retest reliability of acute pain threshold, tolerance, and goodness-of-fit within individuals across ASC sessions. We used intra-class correlation (ICC) to estimate reliability, and used limits of agreement,[12] within subjects coefficients of variation,[11] and Lin's concordance correlation coefficient[43] to measure repeatability or agreement.[10,12] We included all visits in ICC analyses and conducted follow up analyses restricted to the first two visits to evaluate agreement as well as associations with duration between visits (see "Analysis as a function of duration between visits," below).

Because many approaches exist to evaluate reliability and repeatability, and these are implemented differently in different statistical packages, we report results across several approaches for computing reliability and ICC in the statistical software R.[57] We used the same approach for each outcome (ie threshold, tolerance, goodness-of-fit). We focused on two-way agreement using single random raters (ie ICC(2,1)),[39,48,65] since our participants underwent the same procedure in different sessions and the number of visits varied across participants. In the main manuscript, we focus on models that computed ICC in the context of linear mixed models that incorporated fixed effects of visit number, gender, and environment on each outcome measure and treated participant as random (see below, "Effects of environment, gender, and visit number on pain sensitivity"). We used the 'performance' package[46] to compute ICC and used the "repeatability" function from 'rptR'[68] to evaluate repeatability from the same model, as described in Stoffel et al., 2019.[69] We also computed ICC using several additional packages (see Supplementary Methods). Results were highly consistent regardless of analytic approach. All details are reported in Supplemental Materials.

ICC values were interpreted as in Koo & Li,[39] in which values below .5 indicate poor reliability, values between .5 and .75 indicate moderate reliability, values between .75 and .9 indicate good reliability, and excellent reliability is denoted by ICC values above .9. We note that other guidelines have slightly different interpretations, with values above .75 indicating excellent reliability and values between .6 and .74 indicating good reliability.[16]

We evaluated agreement between the first two visits (n = 171) in line with suggestions from Bland and Altman.[12] We computed the limits of agreement using a Bland-Altman plot, which compares the average of two measurements with the difference between the measurements. The mean difference is referred to as the bias, and the 95% confidence interval surrounding the bias provides the limits of agreement. If observations fall within the limits of agreement, repeated measurements show high agreement. We used the function "agree_reps" in the 'SimplyAgree' package[13] to compute the Concordance Correlation Coefficient (CCC), a measure of agreement.[43] We also computed ICC values for the first two visits using the "icc" function in the package 'IRR'[26] and report results for a two-way model of agreement and single raters.

Finally, to provide interpretable values that compare the variation across outcomes, we computed the within-subjects coefficient of variation (WSCV[11]), which evaluates the extent to which within-subjects error varies as a function of overall mean.

Analysis as a function of duration between visits.—A second goal was to test whether there was a significant association between the delay between visits and the consistency of outcome measures. We therefore conducted additional analyses restricted to the first two visits. We measured the number of days between each participant's first two calibration visits and computed correlations with the between-session differences in thresholds, tolerance, and r^2s. Since the duration between the first two visits was not normally distributed (there was a significant skew toward shorter durations), we analyzed the data using Spearman's Rank-Order correlation, implemented in the R package 'stats'.[57] Since the difference of threshold, tolerance, and r^2s between the two visits could be positive or negative, we analyzed both the raw and absolute values of the differences in threshold, tolerance, and r^2s.

Evaluating potential gender differences in reliability.—Our third goal was to evaluate reliability as a function of gender and test whether reliability differed between female participants (n = 99) and male participants (n = 72). We used "bootstrapping" in the R package 'boot'[14,20] to generate a distribution of ICC estimates across random subsamples of participant (1000 bootstrap iterations). On each iteration, a random sample of participants was selected, we calculated ICC separately for male and female participants, and we computed the difference between ICC estimates for male and female participants. We then estimated the 95% confidence interval from the bootstrapped estimated difference, which were normally distributed. If the interval did not contain 0, we concluded that there was a significant difference in reliability as a function of gender. We also use the R package 'ICC'[75] to report ICC estimates and confidence intervals separately for each group.

Controlling for effects of environment, gender, and visit number on pain sensitivity.—In the context of evaluating reliability, we also accounted for specific factors that might have impacted pain sensitivity as measured by the ASC task. We used linear mixed models to test whether outcome measures (threshold, tolerance, and goodness-of-fit) differed by the number of visits, the testing environment (behavioral testing room vs MRI suite), and the participant's gender. Linear mixed models were implemented using "LMER" from the 'lme4' package in R.[9] Each model included fixed effects of visit number, gender, and environment, and we included random intercepts at the level of participant. Models that treated slopes as random for visit number and/or environment did not converge for tolerance or goodness-of-fit, and ICC values were similar for threshold whether or not slopes were modeled as random. We therefore focus on models that treated all factors as fixed and treat rater (ie session) as fixed, corresponding to ICC(3,1).[39,65]

Results

Descriptives.

One hundred seventy-one participants completed the ASC session on more than one visit. Table 1 presents mean pain threshold, tolerance, and r^2 values as a function of visit number for participants who completed multiple visits, and provides estimates for participants who completed only one visit for comparison. Comparisons between participants who completed a single visit and those who completed multiple visits are reported in Supplementary Results. Results including outliers and using nonlinear approaches are consistent and reported in Supplementary Materials. The length of time between the first two visits ranged from 1 to 952 days, with a median of 23 days and an interquartile range of 55.25 days.

Pain thresholds have moderate reliability.

Pain thresholds were moderately reliable across visits (see Fig 2A), regardless of analysis approach (ICC = .658; see Supplementary Table S2 for all approaches and complete statistics). Findings were similar when we included outliers (ICC = .626; see Supplementary

Table S3) or accounted for nonlinear associations between temperature and pain (ICC = .559; see Supplementary Table S4).

When restricted to the first two visits (see Fig 3, top), ICC remained moderate (ICC = .619, CI = [.50, .71]; see Supplementary Table S2), and we observed low to moderate agreement between measures based on Bland-Altman limits of agreement (Fig 3, middle) and the Concordance Correlation Coefficient (CCC = .62, 95% CI [.52, .70]). The within-subjects coefficient of variation (WSCV) was 3.64%.

Pain tolerance is moderately reliable.

Across analytic approaches, we found that pain tolerance was moderately reliable (ICC = .67; see Fig 2B). Reliability was similar when we evaluated tolerance including outliers (linear tolerance: ICC = .624; see Supplementary Table S3) and slightly lower when we accounted for nonlinear associations between temperature and pain (ICC = .431; see Supplementary Table S4).

When restricted to the first two visits (see Fig 3), ICC remained moderate (ICC = .681; CI = [.53, .78]) and we observed low to moderate agreement between measures based on Bland-Altman limits of agreement (Fig 3, middle) and the Concordance Correlation Coefficient (CCC = .68, 95% CI [.62, .73]). The WSCV was 2.23%.

Temperature-pain correlations have low reliability.

In contrast to the moderate reliability of threshold and tolerance measures, the temperature-pain association had low reliability across sessions in all approaches (ICC = .171; see Fig 2C and Supplementary Table S2). Reliability remained low and we observed the same interactions when we included all trials (ICC = .222; see Supplementary Table S3) or accounted for nonlinear associations between temperature and pain (ICC = .181; see Supplementary Table S4).

When we restricted analyses to the first two visits (see Fig 3), ICC remained low whether we included excluded outliers (ICC = .118, CI = [−.108, .253]) or included all trials (ICC= .247, CI = [.104, .381]). Agreement was also low across the two visits regardless of whether outliers were included (with outlier trials: CCC = .246, 95% CI = [.14, .34]; without outlier trials: CCC = .117, CI = [−.12, .25]). The WSCV was 33.76% when including outlier trials and 32.66% when excluding outlier trials, in contrast to the low WSCV values for threshold and tolerance. This indicates that variation across sessions was related to the mean r^2 value, as can be seen in the Bland-Altman plot (Fig 3, middle). Participants who had high reliability in the association between temperature and pain on visit 1 continued to show high reliability during session 2, but individuals who had lower r^2 values on average tended to become less reliable over time.

Duration between visits does not impact reliability.

There was no association between the duration between visits and any of our outcomes, whether measured by actual or absolute differences (all *Ps* > .1, see Fig 3, bottom).

Gender differences in reliability.

We evaluated whether reliability differed as a function of gender for any of the outcome measures. We observed a significant gender difference in reliability for pain tolerance ($95\%CI_{M-F} = [-.29, -.06]$), such that females were more reliable than males. Differences in pain threshold and temperature-pain associations were in the same direction, although not statistically significant (threshold: $95\%CI_{M-F} = [-.19, .07]$; r^2: [−.42, .01]). See Fig 4 and Table 2 for reliability estimated separately for each gender group and comparisons between groups.

Evaluating potential impact of gender, visit number, and environment on pain sensitivity measures.

Pain thresholds did not differ as a function of gender, testing environment (MRI vs behavioral testing room), or the number of visits, nor were there any interactions between these factors (all *Ps* > 01; see Supplementary Table S5). Pain tolerance differed significantly by gender (B = −.332, *P* = .012; see Supplementary Table S5), such that male participants had higher tolerance across sessions (M = 48.31, SD = 1.69) than female participants (M = 47.73, SD = 1.94). Pain tolerance also decreased within individuals across visits (B = −.204, *P* = .011; see Supplementary Table S4), and we observed a significant Environment x Visit number interaction (B = −.251, *P* = .006; see Supplementary Table S5). Post-hoc tests indicate that the interaction was driven by tolerance decreasing over time in the behavioral clinic (B = −.38, *P* < .001), and no effect of visit number in the fMRI center (*P* > .8). Finally, there was no influence of gender on goodness-of-fit between temperature and pain (r^2), but we observed a significant interaction between environment and visit number (B = −.032, *P* = .007; see Supplementary Table S5). Post hoc tests separated by environment indicated that the goodness-of-fit decreased across visits in the outpatient clinic (B = −.03, *P* = .005) whereas there were no associations between goodness-of-fit and time in the fMRI environment (*P* > .2). There were no additional main effects or interactions (all *Ps* > .2).

Discussion

We evaluated the test-retest reliability of suprathreshold thermal pain sensitivity in a large sample of healthy volunteers staircase calibration (ASC).[3,21,49] Despite variations in testing locations, experimenters, and intervals between visits, pain thresholds and tolerance were moderately reliable across visits, indicating that pain sensitivity is relatively stable over time. However, associations between pain and temperature were strikingly inconsistent across visits. Pain tolerance was significantly more reliable in female participants relative to males, and we observed similar patterns for threshold and temperature-pain associations. Finally, we observed no impact of duration between visits on the stability of measurements. Here, we discuss these findings and their implications.

Pain thresholds were moderately reliable over time, regardless of analysis approach, and did not differ as a function of testing environment, number of visits, or gender. This extends previous findings in smaller, clinical samples that indicate moderate reliability of thermal pain thresholds.[15,35,47,53,59] Like pain threshold, pain tolerance was also moderately reliable over time, whether we used linear or non-linear estimation. This supports the use of adaptive calibrations such as the ASC task for QST, and supports their use for longitudinal studies and studies of individual differences, as threshold and tolerance are relatively stable in healthy volunteers.

In contrast to threshold and tolerance, the overall strength of the temperature-pain relationship, as measured by goodness-of-fit (ie r^2), had markedly low reliability across sessions, regardless of analysis approach. This suggests that although pain threshold and tolerance are relatively stable within individuals across visits, individuals' ratings may be more variable between the anchors of pain onset and maximum tolerable pain. Why might we see such dissociations? Inspecting the Bland-Altman plots in Fig 3 provides important insights. First, participants who had high correlations between temperature and pain (ie $r^2 > .8$) had extremely high agreement between sessions, as indexed by the difference between visits falling on the bias line. Second, participants who fall outside the limits of agreement do so systematically, that is they tend to have lower associations on their second visit, which is consistent with the fact that participants were only eligible to complete multiple visits if they exhibited reliable temperature-pain associations on their first visit (ie $r^2 > .4$). Our findings therefore suggest that participants who show high psychophysical accuracy maintain this over time, whereas those who have more variability in the association between temperature and pain show less agreement across visits. Furthermore, participants who completed multiple visits had stronger correlations between temperature and pain on their first visit than participants who completed only one visit, which likely is related to our use of this measure as a screening criterion. Because this measure is highly variable across visits, it may not be appropriate to use goodness-of-fit to determine eligibility. Future work should determine whether there are meaningful individual differences that account for variability across individuals and whether specific contextual factors influence the reliability of the temperature-pain association (leading to variation across visits).

We evaluated reliability in data that were collected over the span of nearly 5 years, including different experimenters, devices, and testing environments. The wide range of intervals between tests allowed us to test whether responses vary as a function of time between visits, for example if recall influences test-retest reliability. In contrast to our hypothesis that pain sensitivity measures would be more stable within individuals when visits were closer in time, we observed no differences as a function of duration between participants' first and second visits in any outcome. This is consistent with previous conclusions, as QST studies with short intervals between visits did not show better reliability than those with longer durations between measurements.[52] We also tested whether pain sensitivity measures showed any consistent effects across participants as a function of visit number. We found no consistent effects of visit number on pain threshold or goodness-of-fit, but pain tolerance decreased across visits. In addition, both pain tolerance and goodness-of-fit showed differences depending on environment: Both measures decreased across time in the outpatient clinic, but not in the fMRI environment. We note that participants' first visits were always in the outpatient clinic and were used for eligibility, whereas there were no such restrictions in subsequent visits, which might contribute to these findings with respect to changes in goodness-of-fit. These changes across visits serve to increase within-person variability and therefore reduce the test-retest reliability of tolerance and goodness-of-fit measurements. Future work should address which factors might lead to systematic changes as a function of experience and testing environment, such as within-person variations in psychological state, such as attention, cooperation, motivation, and anxiety, which have all been shown to influence QST measures,[7] or the psychosocial context surrounding sensory testing, such as coherence with experimenter based on ethnicity or gender,[2,45] as discussed above. A 2020 study found that stability of experimenter is “extremely important” for interpretation of results in studies of test-retest reliability.[44] Thus, reliability of our ASC-derived measures might increase if experimenter and environment were held constant. However, the fact that reliability of pain threshold and tolerance was good across multiple experimenters, different environments, and even up to several years between sessions, indicates that these metrics are quite stable over time.

We also compared reliability as a function of gender. While our findings replicate many studies that find higher pain tolerance in male, relative to female, participants,[6] to our knowledge this is the first study to measure gender differences in the test-retest reliability of pain sensitivity in humans. In contrast to assumptions underlying the exclusion of females in pain research, female participants exhibited *higher* reliability on all pain measures than males, who were more variable across visits. Differences in reliability of pain tolerance were significant, and we observed similar trends for pain threshold and temperature-pain associations. This builds on a previous study that indicated that female participants exhibit better discrimination of thermal pain stimuli relative to male participants across visits, although reliability was not formally evaluated.[22] Our findings refute assumptions that female participants are less reliable due to hormonal fluctuations. The present study was not designed to test whether these gender differences in pain reliability are biological (ie, reflective of sex, rather than gender), learned, or reflective of the testing context (eg gender of the experimenter). Prior work on the stability of pain across visits suggests that neither experimenter gender nor type of rating scale is likely to influence gender differences in pain sensitivity or reliability.[22] Studies that focus on a single experimental session have shown that experimenter gender can contribute to gender differences in pain[2,29,38]; however results are mixed,[22,71] several studies experimentally manipulated features of the experimenter (eg attractiveness[41]), and studies have not measured the reliability of these effects. If experimenter gender impacted responses in our studies, we expect that this contributed to greater variability across sessions, as the experimenter frequently varied across study visits, and we included both male and female experimenters. Future studies should examine gender differences in reliability of other measures of pain sensitivity to evaluate the extent to which biological, contextual, and gender role expectations contribute to reliability. Our findings provide direct evidence to support the critical inclusion of female participants in pain research and refute the groundless assumptions that are still used to justify studies enrolling only male participants.

While our findings of moderate reliability in pain threshold and tolerance are consistent with previous studies, our task has several important differences from standard QST procedures that must be acknowledged. We focused on measures of supra-threshold pain, while most studies have focused primarily on the reliability of thermal detection thresholds and pain thresholds. It is possible that suprathreshold pain ratings are less reliable than discrimination threshold and pain threshold since thresholds correspond to the firing properties of C-mechanoreceptors,[40] whereas suprathreshold pain does not map clearly onto peripheral nociceptor sensitivity and is more likely to depend on central mechanisms. We also used a visual analogue scale (VAS) to obtain pain ratings. The VAS has been widely recognized as feasible and acceptable for clinical evaluations[30] and has been shown to be highly reliable in both literate and illiterate patients.[24] However, the discrete levels of the VAS impose limitations on reporting pain, with only a narrow range of scores that is potentially insensitive to change. Therefore, reliability might differ with other scales (eg those using different anchors[37]; or other pain measures (eg pain biosignatures[33]). We also found that linear models provided better fits than nonlinear models in this rather large dataset, whereas previous work has suggested that the relationship between temperature and pain rating is nonlinear.[67,70] Linear models may have provided better fits for our ASC data because temperatures were selected based on iterative linear regression, which might have encouraged subjects to rate pain linearly. Notably, pain sensitivity outcomes showed similar reliability estimates regardless of whether we used linear or nonlinear models. Future work should determine the factors that influence whether pain is linear or nonlinear with respect to noxious input.

Our study raises important outstanding questions that should be addressed in future work, in addition to those highlighted above. First, we measured ASC task reliability in healthy volunteers between ages 18 to 50, and we need to formally evaluate whether this task or similar adaptive calibrations are reliable in patient populations and older adults. In addition, only participants with reasonable correlations between temperature and pain were invited to participate in multiple visits, so our conclusions are limited to individuals who understood the pain scale and had decent perceptual acuity. Future work should compare reliability of thermal heat pain with other transient pain measures such as shock or pressure, which can also be administered and individually tailored using iterative regression as we do here. Comparing the reliability of pain with other modalities would reveal whether our findings are specific to heat pain or reflect general psychophysical measurement (eg that individuals who show high associations between stimulus and response on an initial visit show higher agreement over time). Only two of our studies presented stimuli of another modality (sugar and salt liquid tastants) during testing, and these studies did not differ substantially from our other tasks (see Supplementary Table S1). We acknowledge that there was heterogeneity in stimulus parameters due to our use of the ASC task to evaluate eligibility and select temperatures for subsequent studies. For instance, most of our studies evaluated pain sensitivity on the volar forearm, with the exception of one procedure. Thermal pain thresholds have also been shown to have different levels of reliability on different areas of the body. We did see slightly lower goodness-of-fit on the study that tested the calf relative to the studies that tested the forearm, however we did not have adequate power to directly compare parameters as a function of skin site. Thus, future work should formally compare reliability on the arm with other skin sites. However, we believe that it is unlikely that the specific variations in stimulus parameters contributed meaningfully to our findings. Across the entire sample of participants who completed this task, variations between study contributed to less than 5% of the variance, whereas variations between individuals contributed to between 20.3% and 91.39% of the variance in outcomes (see Supplementary Results: Assessing variance components).

In conclusion, our study examined the reliability of several measures of suprathreshold pain perception in a large sample of healthy volunteers. Thermal pain threshold and tolerance were moderately reliable within and between individuals and remained relatively stable independent of gender, testing environment, and duration between visits. They may therefore serve as adequate measures to track sensory changes over time as well as to evaluate response to interventions, at least in healthy volunteers. In contrast, goodness-of-fit had low reliability, indicating that it is more sensitive to contextual factors that vary over visits, although individuals with strong associations showed high agreement over time. Importantly, we also showed that female participants are not more variable than males; in fact, females had significantly higher reliability in pain tolerance, and showed similar trends across all measures. This evidence refutes common justifications the exclusion of women in pain research. Our work adds to a body of literature on QST reliability and suggests that different measures of pain sensitivity have different variability across time. Future work on pain and its modulation should continue to understand the contextual factors that contribute to variability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Lauren Banker, Esther Palacios-Barrios, Olga Oretsky, Linda Ellison-Dejewski, Adebisi Ayodele, and the nurses of NIH Outpatient Clinic 4 for their vital role in data collection and clinical support. We also thank Simone Haller for helpful discussions on statistical analysis.

This work was supported in part by the Intramural Research Program of the National Center for Complementary and Integrative Health (ZIA-AT-000030, PI Atlas; NCT02446262 and NCT02707029).

References

1. Arnegard ME, Whitten LA, Hunter C, Clayton JA: Sex as a biological variable: A 5-year progress report and call to action. J Women’s Health 29:858–864, 2020
2. Aslaksen PM, Myrbakk IN, Høifødt RS, Flaten MA: The effect of experimenter gender on autonomic and subjective responses to pain stimuli. Pain 129:260–268, 2007 [PubMed: 17134832]
3. Atlas LY, Bolger N, Lindquist MA, Wager TD: Brain mediators of predictive cue effects on perceived pain. J Neurosci 30:12964–12977, 2010
4. Atlas LY, Lindquist MA, Bolger N, Wager TD: Brain mediators of the effects of noxious heat on pain. Pain 155:1632–1648, 2014 [PubMed: 24845572]
5. Atlas LY, Whittington RA, Lindquist MA, Wielgosz J, Sonty N, Wager TD: Dissociable influences of opiates and expectations on pain. J Neurosci 32:8053–8064, 2012 [PubMed: 22674280]
6. Averbeck B, Seitz L, Kolb FP, Kutz DF: Sex differences in thermal detection and thermal pain threshold and the thermal grill illusion: A psychophysical study in young volunteers. Biol Sex Differ 8:1–13, 2017 [PubMed: 28078076]
7. Backonja MM, Walk D, Edwards RR, Sehgal N, Moeller-Bertram T, Wasan A, Irving G, Argoff C, Wallace M: Quantitative sensory testing in measurement of neuropathic pain phenomena and other sensory abnormalities. Clin J Pain 25:641–647, 2009 [PubMed: 19692807]
8. Bartley EJ, Fillingim RB: Sex differences in pain: A brief review of clinical and experimental findings. Br J Anaesth 111:52–58, 2013 [PubMed: 23794645]
9. Bates D, Mächler M, Bolker B, Walker S: Fitting linear mixed-effects models using lme4. J Stat Softw 67, 2015

10. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 8:307–310, 1986
11. Bland JM, Altman DG: Statistics notes: Measurement error proportional to the mean. BMJ 313:106, 1996 [PubMed: 8688716]
12. Bland JM, Altman DG: Measuring agreement in method comparison studies. Stat Methods Med Res 8:135–160, 1999 [PubMed: 10501650]
13. Caldwell A: SimplyAgree: Flexible and robust agreement and reliability analyses. (R package version 0.0.2), 2021. [Computer software]
14. Canty A, Ripley B: Boot: Bootstrap R (S-Plus) functions. (R package version 1.3–25), 2020. [Computer software]
15. Cathcart S, Pritchard D: Reliability of pain threshold measurement in young adults. J Headache Pain 7:21–26, 2006 [PubMed: 16440140]
16. Cicchetti DV: Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 6:284–290, 1994
17. Clayton JA: Applying the new SABV (sex as a biological variable) policy to research and clinical care. Physiol Behav 187:2–5, 2018 [PubMed: 28823546]
18. Clayton JA, Collins FS: Policy: NIH to balance sex in cell and animal studies. Nature 509:282–283, 2014 [PubMed: 24834516]
19. Davis KD, Aghaeepour N, Ahn AH, Angst MS, Borsook D, Brenton A, Burczynski ME, Crean C, Edwards R, Gaudilliere B, Hergenroeder GW, Iadarola MJ, Iyengar S, Jiang Y, Kong J-T, Mackey S, Saab CY, Sang CN, Scholz J, Pelleymounter MA: Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: Challenges and opportunities. Nat Rev Neurol 16:381–400, 2020 [PubMed: 32541893]
20. Davison AC, Hinkley DV: Bootstrap Methods and Their Application. Cambridge, UK, Cambridge University Press, 1997
21. Dildine TC, Necka EA, Atlas LY: Confidence in subjective pain is predicted by reaction time during decision making. Sci Rep 10:21373, 2020
22. Feine JS, Bushnell CM, Miron D, Duncan GH: Sex differences in the perception of noxious heat stimuli. Pain 44:255–262, 1991 [PubMed: 2052394]
23. Feldhaus MH, Horing B, Sprenger C, Büchel C: Association of nocebo hyperalgesia and basic somatosensory characteristics in a large cohort. Sci Rep 11:762, 2021 [PubMed: 33436821]
24. Ferraz MB, Quaresma MR, Aquino LR, Atra E, Tugwell P, Goldsmith CH: Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. J Rheumatol 17:1022–1024, 1990 [PubMed: 2213777]
25. Fruhstorfer H, Lindblom U, Schmidt WC: Method for quantitative estimation of thermal thresholds in patients. J Neurol Neurosurg Psychiatry 39:1071–1075, 1976 [PubMed: 188989]
26. Gamer M, Lemon J, Fellows I, Singh P, Kendall’s W: Various coefficients of interrater reliability and agreement. R package version 0.84.1, 2019, [Computer software]
27. Geber C, Klein T, Azad S, Birklein F, Gierthmühlen J, Huge V, Lauchart M, Nitzsche D, Stengel M, Valet M, Baron R, Maier C, Tölle T, Treede R-D: Test−retest and interobserver reliability of quantitative sensory testing according to the protocol of the German Research Network on Neuropathic Pain (DFNS): A multi-centre study. Pain 152:548–556, 2011 [PubMed: 21237569]
28. Gehling J, Mainka T, Vollert J, Pogatzki-Zahn EM, Maier C, Enax-Krumova EK: Short-term test-retest-reliability of conditioned pain modulation using the cold-heat-pain method in healthy subjects and its correlation to parameters of standardized quantitative sensory testing. BMC Neurol 16:125, 2016 [PubMed: 27495743]
29. Gijsbers K, Nicholson F: Experimental pain thresholds influenced by sex of experimenter. Percept Mot Skills 101:803–807, 2005 [PubMed: 16491681]
30. González-Fernández M, Ghosh N, Ellison T, McLeod JC, Pelletier CA, Williams K: Moving beyond the limitations of the visual analog scale for measuring pain: Novel use of the general labeled magnitude scale in a clinical setting. Am J Phys Med Rehabil 93:75–81, 2014 [PubMed: 23900013]
31. Grahl A, Onat S, Büchel C: The periaqueductal gray and Bayesian integration in placebo analgesia. ELife 7:e32930, 2018

32. Green B, Cruz A: Warmth-insensitive fields": Evidence of sparse and irregular innervation of human skin by the warmth sense. Somatosens Mot Res 15:269–275, 1998 [PubMed: 9875545]

33. Han X, Ashar YK, Kragel P, Petre B, Schelkun V, Atlas LY, Chang LJ, Jepma M, Koban L, Losin ER, Roy M, Woo CW, Wager TD: Effect sizes and test-retest reliability of the fMRI-based neurologic pain signature. Neuroimage 247:118844, 2022 [PubMed: 34942367]

34. Hansson P, Backonja M, Bouhassira D: Usefulness and limitations of quantitative sensory testing: Clinical and research application in neuropathic pain states. Pain 129:256–259, 2007 [PubMed: 17451879]

35. Heldestad V, Linder J, Sellersjö L, Nordh E: Reproducibility and influence of test modality order on thermal perception and thermal pain thresholds in quantitative sensory testing. Clin Neurophysiol 121:1878–1885, 2010 [PubMed: 20478739]

36. Hilz MJ, Stemper B, Axelrod FB, Kolodny EH, Neundörfer B: Quantitative thermal perception testing in adults. J Clin Neurophysiol 16:462, 1999 [PubMed: 10576229]

37. Hjermstad MJ, Fayers PM, Haugen DF, Caraceni A, Hanks GW, Loge JH, Fainsinger R, Aass N, Kaasa S: Studies comparing numerical rating scales, verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: A systematic literature review. J Pain Symptom Manage 41:1073–1093, 2011 [PubMed: 21621130]

38. Kállai I, Barke A, Voss U: The effects of experimenter characteristics on pain reports in women and men. Pain 112:142–147, 2004 [PubMed: 15494194]

39. Koo TK, Li MY: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15:155–163, 2016 [PubMed: 27330520]

40. LaMotte RH, Campbell JN: Comparison of responses of warm and nociceptive C-fiber afferents in monkey with human judgments of thermal pain. J Neurophysiol 41:509–528, 1978 [PubMed: 418156]

41. Levine FM, Lee De Simone L: The effects of experimenter gender on pain report in male and female subjects. Pain 44:69–72, 1991 [PubMed: 2038491]

42. Leys C, Ley C, Klein O, Bernard P, Licata L: Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J Exp Soc Psychol 49:764–766, 2013

43. Lin LI- K: A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255, 1989 [PubMed: 2720055]

44. Lin W, Zhou F, Yu L, Wan L, Yuan H, Wang K, Svensson P: Quantitative sensory testing of periauricular skin in healthy adults. Sci Rep 10:3728, 2020 [PubMed: 32111937]

45. Losin EAR, Anderson SR, Wager TD: Feelings of clinician-patient similarity and trust influence pain: Evidence from simulated clinical interactions. J Pain 18:787–799, 2017 [PubMed: 28479279]

46. Lüdecke D, Ben-Shachar MS, Patil I, Waggoner P, Makowski D: Performance: An R package for assessment, comparison and testing of statistical models. J Open Source Software 6:3139, 2021

47. Marcuzzi A, Wrigley PJ, Dean CM, Adams R, Hush JM: The long-term reliability of static and dynamic quantitative sensory testing in healthy individuals. Pain 158:1217–1223, 2017 [PubMed: 28328574]

48. McGraw KO, Wong SP: Forming inferences about some intraclass correlation coefficients. Psychol Methods 1:30–46, 1996

49. Mischkowski D, Palacios-Barrios EE, Banker L, Dildine TC, Atlas LY: Pain or nociception? Subjective experience mediates the effects of acute noxious heat on autonomic responses - corrected and republished. Pain 160:1469–1481, 2019 [PubMed: 31107415]

50. Mischkowski D, Stavish CM, Palacios-Barrios EE, Banker LA, Dildine TC, Atlas LY: Dispositional mindfulness and acute heat pain: Comparing stimulus-evoked pain with summary pain assessment. Psychosom Med 83:539–548, 2021 [PubMed: 34213859]

51. Mogil JS, Chanda ML: The case for the inclusion of female subjects in basic science studies of pain. Pain 117:1–5, 2005 [PubMed: 16098670]

52. Moloney NA, Hall TM, Doody CM: Reliability of thermal quantitative sensory testing: A systematic review. J Rehabil Res Dev 49:191, 2012 [PubMed: 22773522]

53. Nothnagel H, Puta C, Lehmann T, Baumbach P, Menard MB, Gabriel B, Gabriel HH, Weiss T, Musial F: How stable are quantitative sensory testing measurements over time? Report on 10-week

reliability and agreement of results in healthy volunteers. J Pain Res 10:2067, 2017 [PubMed: 28919806]

54. Pigg M, Baad-Hansen L, Svensson P, Drangsholt M, List T: Reliability of intraoral quantitative sensory testing (QST). Pain 148:220–226, 2010 [PubMed: 20022428]

55. Pleil JD, Wallace MAG, Stiegel MA, Funk WE: Human biomarker interpretation: The importance of intra-class correlation coefficients (ICC) and their calculations based on mixed models, ANOVA, and variance estimates. J Toxicol Environ Health, Part B 21:161–180, 2018

56. Prendergast BJ, Onishi KG, Zucker I: Female mice liberated for inclusion in neuroscience and biomedical research. Neurosci Biobehav Rev 40:1–5, 2014 [PubMed: 24456941]

57. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2020. Available at: https://www.R-project.org/. Accessed March 14, 2022.

58. Rolke R, Baron R, Maier CA, Tölle TR, Treede RD, Beyer A, Binder A, Birbaumer N, Birklein F, Bötefür IC, Braune S: Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): Standardized protocol and reference values. Pain 123:231–243, 2006 [PubMed: 16697110]

59. Rosner J, Hubli M, Hostettler P, Scheuren PS, Rinert J, Kramer JLK, Hupp M, Curt A, Jutzeler CR: Contact heat evoked potentials: Reliable acquisition from lower extremities. Clin Neurophysiol 129:584–591, 2018 [PubMed: 29414402]

60. Searle SR: Linear Models. New York, NY, Wiley, 1971

61. Shansky RM: Sex differences in behavioral strategies: Avoiding interpretational pitfalls. Curr Opin Neurobiol 49:95–98, 2018 [PubMed: 29414071]

62. Shansky RM: Are hormones a “female problem” for animal research? Science 364:825–826, 2019 [PubMed: 31147505]

63. Shansky RM, Murphy AZ: Considering sex as a biological variable will require a global shift in science culture. Nat Neurosci 24:457–464, 2021 [PubMed: 33649507]

64. Shih YW, Tsai H-Y, Lin F-S, Lin Y-H, Chiang C-Y, Lu Z-L, Tseng M-T: Effects of positive and negative expectations on human pain perception engage separate but interrelated and dependently regulated cerebral mechanisms. J Neurosci 39:1261–1274, 2019 [PubMed: 30552181]

65. Shrout PE, Fleiss JL: Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 86:420–428, 1979 [PubMed: 18839484]

66. Stevens SS: On the psychophysical law. Psychol Rev 64:153–181, 1957 [PubMed: 13441853]

67. Stevens SS: To Honor Fechner and Repeal His Law: A power function, not a log function, describes the operating characteristic of a sensory system. Science 133:80–86, 1961 [PubMed: 17769332]

68. Stoffel MA, Nakagawa S, Schielzeth H: rptR: Repeatability estimation and variance decomposition by generalized linear mixed-effects models. Methods Ecol Evol 8:1639–1644, 2017

69. Stoffel MA, Nakagawa S, & Schielzeth H. An introduction to repeatability estimation with rptR, 2019. Available at: https://cran.r-project.org/web/packages/rptR/vignettes/rptR.html. Accessed March 14, 2022.

70. Sturgeon JA, Tieu MM, Jastrzab LE, McCue R, Gandhi V, Mackey SC: Nonlinear effects of noxious thermal stimulation and working memory demands on subjective pain perception. Pain Med 16:1301–1310, 2015 [PubMed: 25929747]

71. Thorn BE, Clements KL, Ward LC, Dixon KE, Kersh BC, Boothby JL, Chaplin WF: Personality factors in the explanation of sex differences in pain catastrophizing and response to experimental pain. Clin J Pain 20:275–282, 2004 [PubMed: 15322433]

72. Vuilleumier PH, Biurrun Manresa JA, Ghamri Y, Mlekusch S, Siegenthaler A, Arendt-Nielsen L, Curatolo M: Reliability of quantitative sensory tests in a low back pain population. Reg Anesth Pain Med 40:665–673, 2015 [PubMed: 26222349]

73. Wager TD, Atlas LY, Lindquist MA, Roy M, Woo CW, Kross E: An fMRI-based neurologic signature of physical pain. N Engl J Med 368:1388–1397, 2013 [PubMed: 23574118]

74. Wasner GL, Brock JA: Determinants of thermal pain thresholds in normal subjects. Clin Neurophysiol 119:2389–2395, 2008 [PubMed: 18778969]

75. Wolak ME, Fairbairn DJ, Paulsen YR: Guidelines for estimating repeatability. Methods Ecol Evol 3:129–137, 2012

76. Yarnitsky D, Granot M: Chapter 27 Quantitative sensory testing. Handbook of Clinical Neurology, 81. Elsevier, 2006, pp 397–409 [PubMed: 18808849]

77. Yarnitsky D, Sprecher E, Zaslansky R, Hemli JA: Heat pain thresholds: Normative data and repeatability. Pain 60:329–332, 1995 [PubMed: 7596629]

78. Zhang S, Yoshida W, Mano H, Yanagisawa T, Mancini F, Shibata K, Kawato M, Seymour B: Pain control by co-adaptive learning in a brain-machine interface. Curr Biol 30:3935–3944, 2020. e7 [PubMed: 32795441]

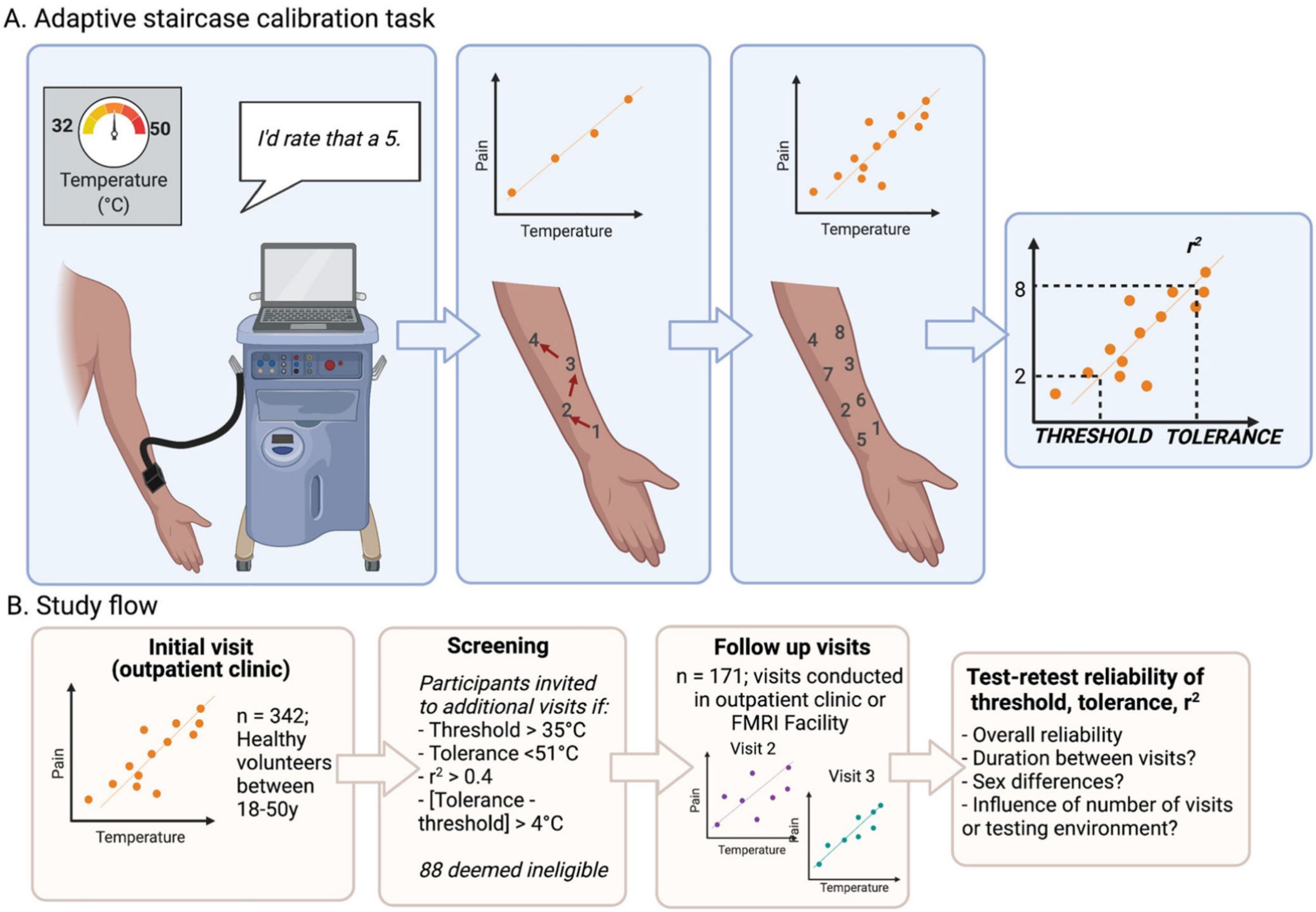
A. Adaptive staircase calibration task
32
50
Temperature (°C)
I'd rate that a 5.
Pain
Temperature
Pain
Temperature
4
3
2
1
4
8
7
3
2
6
5
1
r²
8
2
THRESHOLD
TOLERANCE
B. Study flow
Initial visit (outpatient clinic)
Pain
Temperature
n = 342; Healthy volunteers between 18-50y
Screening
Participants invited to additional visits if:
- Threshold > 35°C
- Tolerance <51°C
- r² > 0.4
- [Tolerance - threshold] > 4°C
88 deemed ineligible
Follow up visits
n = 171; visits conducted in outpatient clinic or FMRI Facility
Visit 2
Pain
Temperature
Visit 3
Pain
Temperature
Test-retest reliability of threshold, tolerance, r²
- Overall reliability
- Duration between visits?
- Sex differences?
- Influence of number of visits or testing environment?

Figure 1.

Adaptive staircase calibration procedure and study flow. Participants underwent an adaptive staircase calibration procedure on each visit. (**A**) Adaptive staircase calibration task. Noxious heat was delivered using a thermode and participants provided pain ratings using a 0 to 10 visual analogue scale after each temperature (left). We iteratively fit a linear regression between temperature and pain and rotated through eight skin sites (middle). After three trials on each skin site (see Methods), we determined the participant's pain threshold (ie the temperature corresponding to a pain rating of 2), tolerance (ie the temperature corresponding to a pain rating of 8), and used r^2 as a measure of goodness-of-fit. (**B**) Study flow. 342 healthy volunteers underwent an initial ASC visit in an outpatient clinic to evaluate pain sensitivity and eligibility for subsequent experiments. Participants who met eligibility criteria were invited to subsequent visits; 88 were deemed ineligible. One hundred seventy-one participants completed multiple visits conducted either in the outpatient clinic or in a suite adjacent to the MRI scanner. We examined the reliability of threshold, tolerance, and goodness-of-fit across study visits, and explored the impact of duration between visits, participant gender, number of visits, and testing environment. This figure was created with BioRender.com. Abbreviations: ASC, adaptive staircase calibration; MRI, magnetic resonance imaging.

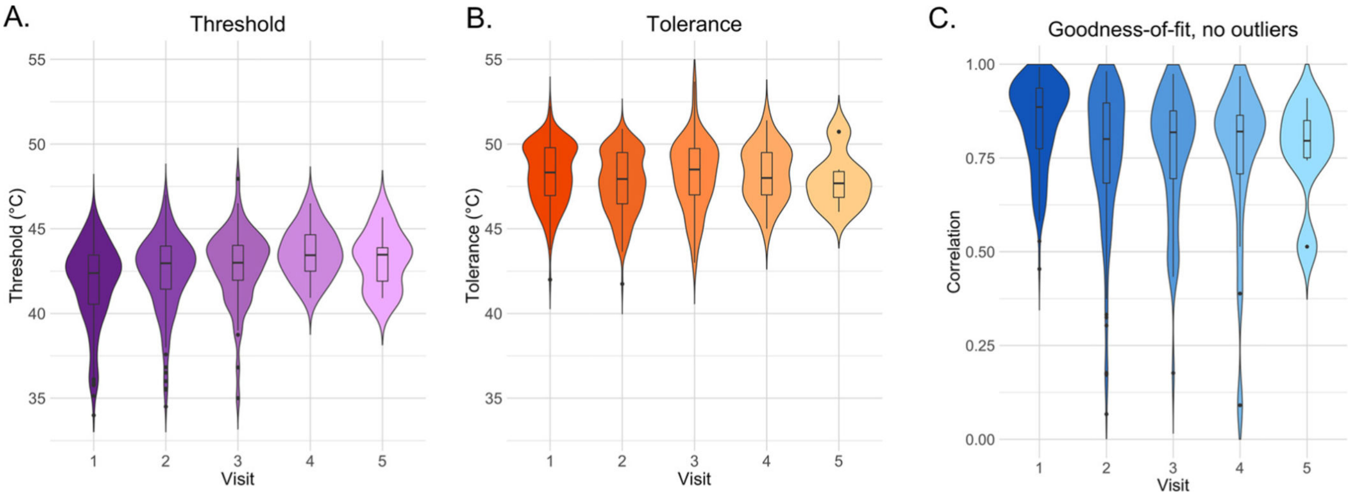

Figure 2.

Threshold, tolerance, and goodness-of-fit by visit for participants who completed multiple visits. Violin plots and bar graphs depict pain threshold (left, purple), tolerance (middle, orange), and correlation between temperature and rating (right, blue) as derived by an adaptive staircase calibration procedure on each visit. Figures depict data from subjects who completed multiple calibrations. Hue darkness reflects the number of participants included in each type of visit: see Table 1 for exact numbers. Goodness-of-fit reflects the correlation between temperature and reported pain with outlier trials excluded.

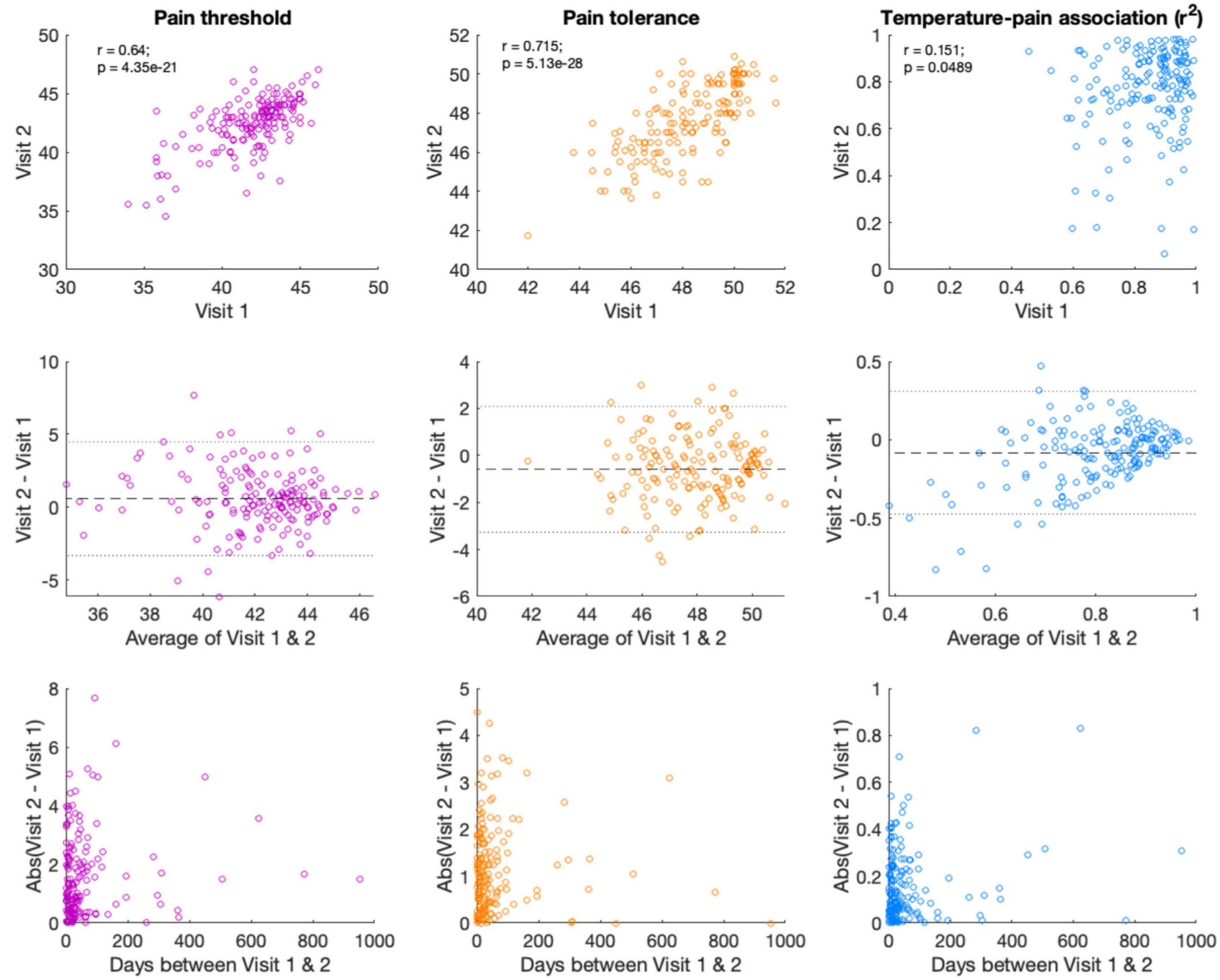

Figure 3.

Associations between temperature, temperature, and goodness-of-fit on visit 1 and visit 2. We evaluated reliability for all measures across the first two visits in participants who completed multiple visits (n = 171). Top: Across the first two visits, we observed high correlations in pain thresholds (left, purple) and pain tolerance (middle, orange) but low correlations in the goodness-of-fit between temperature and subjective pain (right, blue). All estimates are based on analyses excluding outlier trials. Middle: Bland-Altman figures[12] indicate that there was low agreement for the goodness-of-fit measure, although all outcomes included some estimates outside of the limits of agreement. Y-axis depicts the difference between the two visits, and x-axis depicts the average of the two visits. Dashed line depicts the mean of the differences, with dotted lines representing the 95% confidence interval (+/− 2SD). Bottom: Associations did not differ as a function of duration between visits, whether we measured the absolute value of the difference in outcomes across visits (pictured) or the signed difference (all *P*'s > .1).

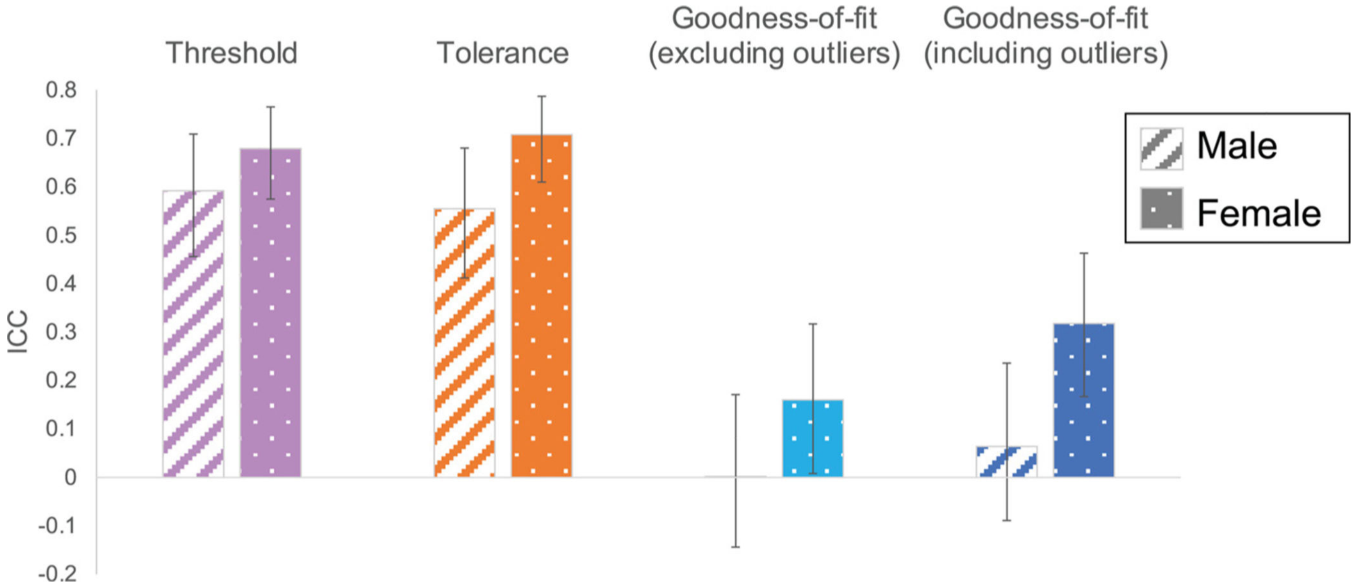

Figure 4.

ICC by gender. We separately evaluated ICC as a function of participant gender and compared groups to determine whether males (left bars) and females (right bars) differed in the reliability of pain threshold, tolerance, or goodness-of-fit (based on analyses excluding outliers; see Supplementary Materials for analyses including outliers and based on nonlinear estimates). Error bars depict 95% confidence intervals. Female participants displayed higher reliability than males across visits in all measures of pain sensitivity, and reliability differed significantly for pain tolerance based on bootstrapped estimation ($95\%CI_{M\text{-}F} = [-.29, -.06]$). See Table 2 for exact values. Abbreviation: ICC, intra-class correlation.

Table 1.

Participation and Pain Sensitivity by Visit Number[*]

VISIT NUMBER		N	THRESHOLD	TOLERANCE	GOODNESS OF FIT (R^2 BETWEEN TEMPERATURE AND PAIN), ALL TRIALS	GOODNESS OF FIT (R^2 BETWEEN TEMPERATURE AND PAIN), EXCLUDING OUTLIERS
1	All participants	342	M = 41.92, SD = 3.16	M = 48.59, SD = 3.81	M = .67, SD = .18	M = .78, SD = .19
	Participants with multiple visits	171	M = 41.80, SD = 2.43	M = 48.23, SD = 1.78	M = .72, SD = .13	M = .85, SD = .11
	Participants with one visit	171	M = 42.02, SD = 3.73	M = 48.95, SD = 5.07	M = .63, SD = .21	M = .72, SD = .22
2		171	M = 42.40, SD = 2.25	M = 47.66, SD = 1.83	M = .68, SD = .18	M = .76, SD = .18
3		61	M = 42.81, SD = 2.21	M = 48.17, SD = 2.09	M = .67, SD = .18	M = .77, SD = .16
4		21	M = 43.57, SD = 1.47	M = 48.24, SD = 1.62	M = .70, SD = .18	M = .76, SD = .21
5		6	M = 43.16, SD = 1.76	M = 47.88, SD = 1.66	M = .67, SD = .09	M = .77, SD = .14

[*] This table reports threshold, tolerance, and goodness-of-fit as a function of visit number, based on an adaptive staircase calibration administered on each visit. The column labeled N reports the number of participants at each visit, and the four rightmost columns report mean and standard deviation for each measure.

Table 2.

Gender Differences in Reliability of Pain Sensitivity Measures*

	THRESHOLD		TOLERANCE		GOODNESS-OF-FIT (ALL TRIALS)		GOODNESS-OF-FIT (EXCLUDING OUTLIER TRIALS)	
	ICC	CI	ICC	CI	ICC	CI	ICC	CI
Male	.592	[.46, .71]	.555	[.41, .68]	.064	[−.09, .24]	.002	[−.14, .17]
Female	.679	[.58, .77]	.708	[.61, .79]	.318	[.17, .46]	.160	[.01, .32]
Male - Female		[−.19, .07]		[−.29, −.06]		[−.47, .01]		[−.34, .17]

[*] This table depicts intraclass correlations (ICC) and confidence intervals (CI) separately for males (n = 72) and females (n = 99). ICC and CI values were determined using “ICCest” in the ‘ICC’ package,[75] with confidence intervals based on Searle[60].