Test-retest reliability of an adaptive thermal pain calibration procedure in healthy volunteers

Carolyn Amir[1]*, Margaret Rose-McCandlish[1]*, Rachel Weger[1]*, Troy Dildine[1,4], Dominik Mischkowski[5], Elizabeth Necka[1,6], In-seon Lee[7, 8], Tor D. Wager[9], Daniel S. Pine[2], and Lauren Y. Atlas[123**]

1. National Center for Complementary and Integrative Health, National Institutes of Health, Bethesda, MD
2. National Institute of Mental Health, National Institutes of Health, Bethesda, MD
3. National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD
4. Clinical Neuroscience Section, Karolinska Institutet, Solna, Sweden
5. Ohio University, Athens, OH, USA
6. National Institute on Aging, National Institutes of Health, Bethesda, MD
7. College of Korean Medicine, Kyung Hee University, Seoul, Republic of Korea
8. Acupuncture & Meridian Science Research Center, Kyung Hee University, Seoul, Republic of Korea
9. Dartmouth College, Hanover, NH, USA

* These authors contributed equally to this work.

** Correspondence should be addressed to:

Dr. Lauren Y. Atlas, PhD
National Institutes of Health (NCCIH, NIMH, NIDA)
10 Center Drive
Bethesda, MD 20892
301-827-0214
lauren.atlas@nih.gov

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 0-10 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 [ICC]=0.66 and 0.67, respectively; p<.001), whereas temperature-pain correlations had low reliability (ICC=0.23). 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, humans, QST, healthy volunteers, intraclass correlations

Quantitative sensory testing (QST) is a valuable psychophysical tool for pain assessment in healthy volunteers (Geber et al., 2011) and in clinical populations (Backonja et al., 2009). QST complements bedside examinations by using standardized procedures to evaluate sensory thresholds and suprathreshold pain perception (Hansson et al., 2007). This permits comparisons with normative data (Rolke et al., 2006) 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 for pain assessment and diagnosis. Reliability is particularly important when evaluating pain biomarkers or signatures (Pleil et al., 2018), as subjective pain is the gold standard against which any pain biomarker is compared (Davis et al., 2020). If QST metrics are sensitive to trait-level factors (i.e. they show larger differences between individuals than within individuals over time), they may be suitable for biomarkers designed to predict diagnosis, risk, or clinical outcomes (Kragel et al., 2021). Alternatively, if metrics are more sensitive to state level factors (i.e. variations within participants across sessions but high consistency within sessions), they point to variations in psychological state, context, or biological processes that may be relevant for predicting immediate treatment response. Both types of biomarkers are suitable for pain and other clinical outcomes (Kragel et al., 2021), as pain is influenced by both state and trait level factors (Davis & Cheng, 2019).

Most QST reliability studies focus on warm and cool detection thresholds, but a few have focused on pain thresholds, i.e. the temperature at which a stimulus is labeled as painful, which corresponds to the activation of peripheral nociceptive C fibers (LaMotte & Campbell, 1978). Test-retest reliability estimates of heat pain threshold range from poor (Yarnitsky et al., 1995) to excellent (Pigg et al., 2010; Wasner & Brock, 2008); however, these studies vary in quality as well as methodology (Moloney et al., 2012). Sample sizes range from 10 participants without pain (Felix & Widerström-Noga, 2009) to 72 healthy volunteers (Yarnitsky et al., 1995), and duration between sessions ranges from 1 day to three weeks (Wasner & Brock, 2008). Our goal was to build on this work by focusing on the reliability of three measures of suprathreshold pain: 1) pain threshold; 2) pain tolerance, or the maximum pain an individual is willing to tolerate; and 3) temperature sensitivity, as measured by the strength of the association between temperature and subjective pain. To address limitations of previous work, we tested reliability in a large sample (n = 171), measured whether reliability varies systematically as a function of duration between visits, and compared multiple analytic approaches to estimate reliability, since previous studies vary in approaches (Moloney et al., 2012) and there are disagreements over the best statistical method to assess test-retest reliability (Berchtold, 2016; Bruton et al., 2000; Qin et al., 2019). For instance, the common intraclass correlation (ICC) tests relative reliability, i.e. whether the overall variation across and within subjects is retained across measurements, whereas repeatability or agreement evaluate the likelihood that an individual will obtain the same score on repeated visits (Bland & Altman, 1986, 1999).

We evaluated the test-retest reliability of a thermal pain QST procedure called the adaptive staircase calibration (ASC; (Atlas et al., 2010; Dildine et al., 2020; Mischkowski et al., 2019)), in which participants provide pain ratings after heat offset using a visual analogue scale (VAS). In contrast to standard QST procedures which rely on the method of limits or method of levels to identify detection thresholds, adaptive staircase calibrations are often used to select suprathreshold stimuli for use in subsequent experiments (Atlas et al., 2010; Feldhaus et al., 2021; Grahl et al., 2018; Shih et al., 2019; Zhang et al., 2020), as they ensure participants can tolerate repeated stimulations at a given intensity. 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 (Wager 2013). The NPS has high reliability both within and across participants (Han et al., 2021); how reliable are the subjective pain measures that were used to train the NPS?

171 healthy volunteers completed the ASC task on multiple visits and we measured the test-retest reliability and agreement of pain threshold, pain tolerance, and the association between temperature and pain. We also evaluated whether reliability is influenced by time

between measurements or testing environment. We hypothesized that pain measures would be more stable when visits were closer in time. Finally we compared reliability between male and female participants to explore whether reliability is lower in female participants. This would be expected if hormonal fluctuations in women are associated with greater variability across sessions, a rationale that is often used to justify excluding women and female animals from pain studies (Shansky, 2018, 2019; Shansky & Murphy, 2021).

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 six 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 neither recreational drugs in the previous month nor taken any pain relievers within 5 half-lives.

342 participants provided informed consent under an NIH IRB-approved protocol (16-AT-0077 and/or 15-AT-0132) and completed at least one visit, during which they underwent an adaptive staircase calibration (ASC) with noxious heat. The ASC task established pain sensitivity and was used to determine eligibility for subsequent experiments. Subsets of these data were included in previous papers on autonomic responses to heat (Mischkowski et al., 2019),

relationships with trait mindfulness (Mischkowski et al., 2021) and confidence in subjective pain (Dildine et al., 2020). In the current paper, we focus on all participants who completed multiple heat calibrations administered on different days (n = 171). Participants had an average age of 28.60 ± 7.83 years; 99 were female (57.90%) and 72 were male (42.11%). 86 described themselves as White (50.29%), 39 as Black/African American (22.81%), 27 as Asian (15.79%), 13 as Hispanic/Latino (7.60%), 4 as more than one race (2.34%), 1 as Native Hawaiian (0.58%), and 14 other or unknown (8.89%).

Stimuli and Apparatus

Heat stimulation was administered to the left volar forearm via a 16x16mm square ATS (Advanced Thermal Stimulator) thermode and controlled with a Pathway Pain and Sensory Evaluation System (Medoc Advanced Medical Systems Ltd, Ramat Yishay, Israel). The thermode was strapped to the arm with Velcro and moved by the experimenter after each stimulation. 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.

General Procedures

General procedures have been covered in detail in previous publications (Dildine et al., 2020; Mischkowski et al., 2019). 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 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 medical ineligibility. The ASC was used to determine pain sensitivity and eligibility for subsequent testing in pain modulation experiments.

We describe eligibility criteria below. The ASC procedure was repeated on all visits. The present analysis focuses only on participants who completed the ASC task on more than one visit (*n* = 171).

Adaptive Staircase Calibration Procedure

The ASC procedure has been described in detail in previous work (Atlas et al., 2010, 2012, 2014; Dildine et al., 2020; Mischkowski et al., 2019, 2021). All participants underwent this task during an initial screening visit, in which they received 24 trials of heat on the left volar forearm (see Figure 1). In brief, we applied 3 rounds of noxious heat to 8 skin sites, with temperatures determined through an iterative regression procedure. Stimuli were either 8s or 10s in duration (see Table 1) including 3s ramping between baseline and target temperature. After the offset of each heat stimulus, participants rated it on a 0-10 visual analog scale (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. (Dildine et al., 2020). 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. Outlier trials were defined as ratings that exceeded 2.5 times the median absolute deviation (Leys et al., 2013) and were excluded from the linear regression.

Trial order was the same across participants to ensure that each site received stimuli at each predicted pain intensity (i.e. 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 0.5°C, unless that temperature or a lower temperature had already been rated as intolerable (i.e. > 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 0.5°C or 1°C below the ASC-predicted temperature. The regression was based on the applied temperature rather than the predicted temperature.

Participants completed the ASC task during an initial visit and then completed the task on each subsequent visit if they were eligible to continue as described in the following section. 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 (see Table 1). In particular, heat stimuli lasted either 8 or 10 seconds, 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 taste stimuli with a similar procedure. These participants also rated pain unpleasantness as well as intensity. We focus on pain intensity ratings alone.

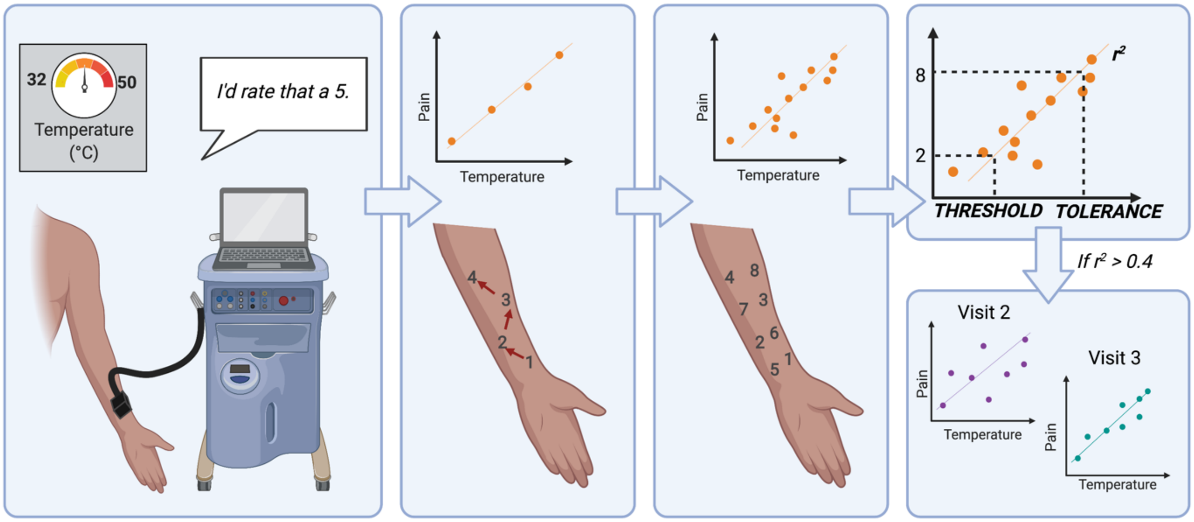

Figure 1. Adaptive staircase calibration procedure. Participants underwent an adaptive staircase calibration procedure on each visit. Noxious heat was delivered using a thermode and participants provided pain ratings using a 0-10 visual analogue scale after each temperature (left). We iteratively fit 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 (i.e. the temperature corresponding to a pain rating of 2), tolerance (i.e. the temperature corresponding to a pain rating of 8), and used r^2 as a measure of goodness-of-fit. Participants who had r^2 greater than 0.4 were invited to subsequent visits, and we examined the reliability of threshold, tolerance, and reliability across visits. This figure was created with BioRender.com.

Threshold, tolerance, and goodness-of-fit estimation. For each individual and each visit, 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. Participants were ineligible for experimental tasks and subsequent visits if their ratings were not ordinally consistent with temperature and they had an r^2 value of less than 0.4 (based on calculation without outliers; n = 16 ineligible). Participants were also excluded from follow up studies if they had a pain threshold below 36°C (*n* = 11), pain tolerance above 50°C (*n* = 43), or a difference of less than 4°C between threshold and tolerance (*n* = 36).

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. Our main analyses focus on estimates that exclude outlier trials to account for warmth-insensitive skin fields in the forearm (Green & Cruz, 1998), and we evaluate goodness-of-fit both with and without outliers. Because mean goodness-of-fit was highest using the ASC task excluding outlier trials, we focus on these results in the main manuscript unless stated otherwise. 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 (Stevens, 1957, 1961). See Supplementary Methods for complete details.

Evaluating reliability, repeatability, and agreement. We were interested in quantifying test-retest reliability in 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 (Bland & Altman, 1999), within subjects coefficients of variation (Bland & Altman, 1996), and Lin's concordance correlation coefficient (L. I.-K. Lin, 1989) to measure repeatability or agreement (Bland & Altman, 1986, 1999). For all participants who completed multiple ASC visits (*n* = 171, see Table 2), we included all visits in ICC analyses. We also 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 (R Core Team, 2020). We used the same approach for each outcome (i.e. threshold, tolerance, goodness-of-fit). We focused on two-way agreement using single random raters (i.e. ICC(2,1); (Koo & Li, 2016; McGraw & Wong, 1996; Shrout & Fleiss, 1979)), 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, sex, and environment on each outcome measure and treated participant as random (see below, “Effects of environment, sex, and visit number on pain sensitivity”). We used the performance package (Lüdecke et al., 2021) to compute ICC and used the “repeatability” function from rptR (Stoffel et al., 2017) to evaluate repeatability from the same model, as described in Stoffel et al, 2019 (Stoffel et al., 2019). 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 (Koo & Li, 2016), in which values below 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and excellent reliability is denoted by ICC values above 0.9. We note that other guidelines have slightly different interpretations, with values above 0.75 indicating excellent reliability and values between 0.6 and 0.74 indicating good reliability (Cicchetti, 1994).

We evaluated agreement between the first two visits (n = 171) in line with suggestions from Bland and Altman (Bland & Altman, 1999). 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 (Caldwell, 2021) to compute the Concordance Correlation Coefficient (CCC), a measure of agreement (L. I.-K. Lin, 1989). We also computed ICC values for the first two visits using the “icc” function in the package ‘IRR’ (Gamer et al., 2012) 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 (Bland & Altman, 1996), which evaluates the extent to which within-subjects error varies as a function of overall mean.

Effects of environment, sex, and visit number on pain sensitivity.

We were interested in testing whether specific factors impacted pain sensitivity, in addition to looking at overall reliability. 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 versus MRI suite), and the participant’s sex. Linear mixed models were implemented using LMER from the lme4 package in R (Bates et al., 2015). Each model included fixed effects of visit number, sex, 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 (i.e. session) as fixed, corresponding to ICC(3,1) (Koo & Li, 2016; Shrout & Fleiss, 1979).

We were also interested in evaluating reliability as a function of sex and testing whether reliability differed between female participants (*n* = 99) and male participants (*n* = 72). We used bootstrapping in the R package boot (Canty & Ripley, 2020; Davison & Hinkley, 1997) 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 sex. We also use the R package ICC (Wolak et al., 2012) to report ICC estimates and confidence intervals separately for each group.

Analysis as a function of duration between visits.

In addition to analyses across all visits, we conducted additional analyses restricted to the first two visits to test whether there was a significant association between the delay between visits and the consistency of outcome measures. Specifically, for the subjects who completed multiple calibrations (*n*= 171), we measured the time between their first two calibration visits in days and computed correlations with the between-session differences in thresholds, tolerance, and *r2*s. We focused on the first two visits for better correspondence with previous work (Moloney et al., 2012) and so that we had equal numbers of observations across subjects. 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' (R Core Team, 2020). Since the difference of threshold, tolerance, and *r2*s between the two visits could be positive or negative, we analyzed both the raw and absolute values of the differences in threshold, tolerance, and *r2*s.

Results

Descriptives. 342 participants completed the ASC task on an initial visit, and 171 participants completed more than one ASC session (see Table 2). For subjects who completed more than one ASC task, 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. Across all subjects, the mean pain threshold on the initial visit was 41.94°C (*SD* = 3.15), the mean pain tolerance was 48.59 (*SD* =

3.82), and the mean r^2 was 0.78 (*SD* = 0.19), or 0.67 (*SD* = 0.18) when outliers were included (see Table 2). Results including outliers and using nonlinear approaches are consistent and reported in Supplementary Materials.

Pain thresholds have moderate to good reliability. Pain thresholds were moderately reliable across visits (see Figure 2A), regardless of analysis approach (ICC = 0.658; see Supplementary Table S1 for all approaches and complete statistics). Thresholds did not differ as a function of sex, 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 S4). Findings were similar when we included outliers (ICC = 0.626; see Supplementary Table S2) or accounted for nonlinear associations between temperature and pain (ICC = 0.559; see Supplementary Table S3).

When restricted to the first two visits (see Figure 3, top), ICC remained moderate (ICC = 0.619, CI = [.50, .71]; see Supplementary Table S1), and we observed low to moderate agreement between measures based on Bland-Altman limits of agreement (Figure 3, middle) and the Concordance Correlation Coefficient (CCC = 0.62, 95% C.I. [0.52, 0.70]). The within-subjects coefficient of variation was 3.64%. There was no association between actual or absolute difference in threshold the duration between visits (all *ps*> 0.1, see Figure 3, bottom).

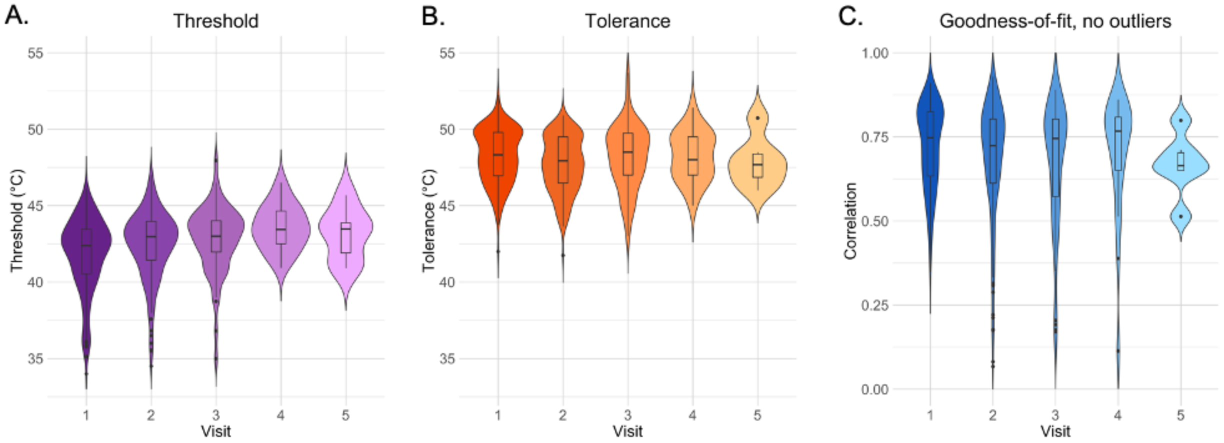

Figure 2. Threshold, tolerance, and goodness-of-fit by visit for participants who completed multiple visits. Violin plots and bar graphs depict pain threshold (left, purple), tolerance (middle, orange), and correlation between temperature and rating (right, blue) as derived by an adaptive staircase calibration 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 2 for exact numbers. Goodness-of-fit reflects the correlation between temperature and reported pain with outlier trials excluded.

Pain tolerance is moderately reliable and differs by sex. Across analytic approaches, we found that pain tolerance was moderately reliable (ICC = 0.67; see Figure 2B). Pain tolerance differed significantly by sex, such that males had higher tolerance (B = -0.332, *p* = 0.012; see Supplementary Table S4), and decreased within individuals across visits (B = -0.204, *p* = 0.011; see Supplementary Table S4). We also observed a significant Environment x Visit number interaction (B = -0.251, *p* = 0.006; see Supplementary Table S4), driven by tolerance decreasing over time in the behavioral clinic (B = -0.38, *p* < 0.001), and no effect of visit number in the fMRI center (*p* > 0.8). Reliability was similar when we evaluated tolerance including outliers (linear tolerance: ICC = 0.624; see Supplementary Table S2) and slightly lower when we accounted for nonlinear associations between temperature and pain (ICC = 0.431; see Supplementary Table S3).

When restricted to the first two visits (see Figure 3), ICC remained moderate (ICC = 0.683; CI = [.59, .76]) and we observed low to moderate agreement between measures based on Bland-Altman limits of agreement (Figure 3, middle) and the Concordance Correlation Coefficient (CCC = 0.68, 95% C.I. [0.62, 0.74]). The within-subjects coefficient of variation was 2.21%. We observed no actual or absolute difference in tolerance as a function of duration between visits (all *ps* > 0.1, see Figure 3, bottom).

Temperature-pain correlations have low reliability and effects across time differ by environment. In contrast to the moderate reliability of threshold and tolerance measures, the temperature-pain association had low reliability across sessions in all approaches, whether outliers were included when computing r^2 (ICC = 0.171; see Supplementary Table S1), or when outliers were excluded (ICC = 0.225; see Figure 2C and Supplementary Table S1). We focus on the goodness-of-fit measure excluding outliers (see Figure 2C and Supplementary Table S4). There was a significant interaction between environment and visit number (B = -0.032, *p* = .007; see Supplementary Table S4). Post hoc tests separated by environment indicated that the goodness-of-fit between temperature and pain decreased across visits in the outpatient clinic (B = -0.03, *p* = .005) whereas there were no associations between goodness-of-fit and time in the fMRI environment (*p* > 0.2). There were no additional main effects or interactions (all *ps* > 0.2). Reliability remained low and we observed the same interactions when we included all trials (ICC = 0.221; see Supplementary Table S2) or accounted for nonlinear associations between temperature and pain (ICC = 0.181; see Supplementary Table S3).

When we restricted analyses to the first two visits (see Figure 3), ICC remained low whether we included all trials (ICC = 0.118, CI = [-.108, .253]) or excluded outliers (ICC= 0.247, CI = [.104, .381]; see Supplementary Table S1). Agreement was also low (CCC = 0.287, 95% C.I. [0.1788, 0.3884]) across the two visits. The within-subjects coefficient of variation was 29.58%, 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 (figure 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. We observed no actual or absolute difference as a function of duration between visits (all p's > 0.4, see Figure 3, bottom).

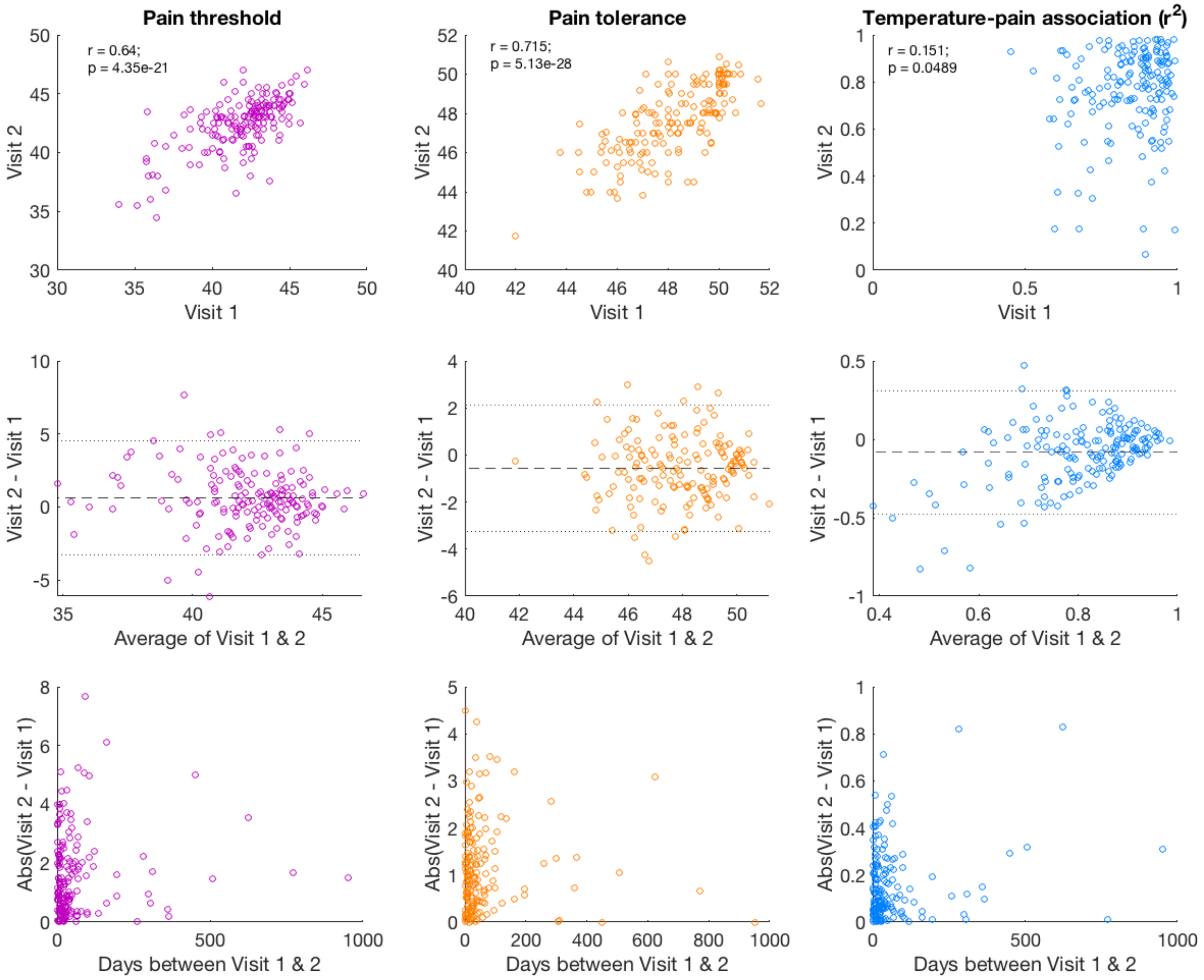

Figure 3. Associations between temperature, temperature, and goodness-of-fit on visit 1 and visit 2. We evaluated reliability for all measures across the first two visits in participants who completed multiple visits (n = 171). *Top:* Across the first two visits, we observed high correlations in pain 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 (Bland & Altman, 1999) 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 > 0.1).

Sex differences. Finally, we evaluated whether reliability differed as a function of sex for any of the outcome measures. We observed a significant sex difference in reliability for pain tolerance (95%CI$_{M-F}$ = [-0.29,-0.06]), 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$_{M-F}$ = [-0.19,0.07]; r2 trim: [-0.42,0.01]). See Figure 4 and Table 3 for reliability estimated separately for each group.

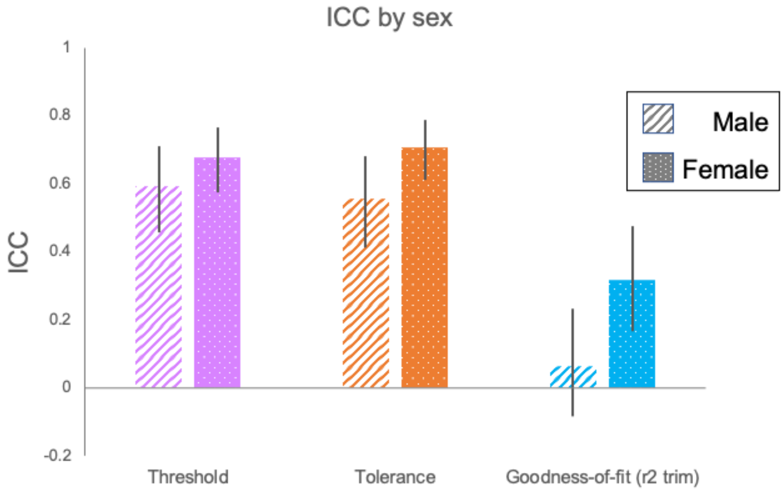

Figure 4. ICC by sex. We separately evaluated ICC as a function of participant sex and compared groups to determine whether males (left bars) and females (right bars) differed in the reliability of pain threshold, tolerance, or goodness-of-fit (based on analyses excluding outliers; see Supplementary

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-F} = [-0.29,-0.06]). See Table 3 for exact values.

Discussion

Quantitative sensory testing (QST) is a critical tool in that facilitates the experimental study of pain by allowing researchers to study pain and its modulation across individuals and within individuals over time. Yet the reliability of metrics derived from QST vary widely (Moloney et al., 2012). We evaluated the test-retest reliability of suprathreshold pain sensitivity using an adaptive heat pain calibration task, the adaptive staircase calibration (ASC; Mischkowski 2019, Atlas 2010, Dildine 2020). 171 healthy volunteers completed the ASC task multiple times with varying locations, investigators, and intervals between visits. Despite these variations, 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, in contrast to assumptions of higher variability in females that have historically been used to exclude female participants from research. Here, we discuss these findings and their implications for QST and studies of pain in general.

Previous studies of QST reliability have focused in large part on thermal detection thresholds rather than pain thresholds; as such, the reliability of thermal pain thresholds is less well established (Moloney 2012). Our analyses revealed that pain thresholds were moderately reliable over time, regardless of analysis approach, and did not differ as a function of sex, testing environment, or number of visits. This extends previous findings in smaller, clinical samples that indicate moderate reliability of thermal pain thresholds (Cathcart 2006; Cruz-Almeida 2015; Heldestad 2010; Marcuzzia 2017; Nothnagel 2017; Rosner 2018).

Our study also expands previous literature by examining changes not just in pain threshold, but also two measures of supra-threshold pain sensitivity: pain tolerance and the goodness-of-fit of the relationship between temperature and pain. Like pain threshold, pain tolerance was also moderately reliable over time, whether we used linear or non-linear estimation. Interestingly, pain tolerance also decreased across visits, particularly when participants were evaluated in an outpatient behavioral clinic, relative to an fMRI scanning environment. In line with previous findings, pain tolerance differed significantly by sex, such that males exhibited higher pain tolerance than females across sessions (Averbeck 2017). However, we found that males did not have higher reliability than female participants; in fact, female participants showed significantly higher reliability than males, as discussed below.

In contrast to the threshold and tolerance measures, the overall strength of the temperature-pain relationship, as measured by goodness-of-fit (i.e. r^2), had markedly low reliability across sessions, regardless of analysis approach or whether we used linear or nonlinear fits. 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 Figure 3 provides important insights. First, participants who had high correlations between temperature and pain (i.e. $r^2 > 0.8$) had extremely high agreement between sessions, as indexed by the difference between visits falling on the bias line. Second, we see that participants who fall outside the limits of agreement do so systematically, i.e. 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 (i.e. $r^2 > 0.4$). 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. Future work should determine whether there are meaningful individual differences that account for this variability across individuals and whether specific contextual factors influence the reliability of the temperature-pain association (leading to variation across visits). For instance, there was a significant interaction between environment and visit number for temperature-pain correlations, such that correlations decreased across visits in the outpatient clinic but not 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. However, participants might also differ in motivation or engagement over time. These alternatives should be investigated in future work.

We also compared reliability as a function of sex in light of the need for greater emphasis on the measurement of sex as a biological variable (Arnegard et al., 2020; Clayton, 2018; Clayton & Collins, 2014). Many preclinical and clinical studies restrict experiments to males due to the implicit assumption that females may be more variable than males due to hormonal fluctuations (Shansky, 2019), even though this assumption has been shown to be erroneous in rodent studies in general (Prendergast et al., 2014) and in rodent models of pain (Mogil & Chanda, 2005). To our knowledge this is the first study to measure sex differences in the test-retest reliability of pain sensitivity in humans. In contrast to implicit assumptions, female participants exhibited *higher* reliability on all pain measures than males, who were more variable across visits. Differences in 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 (Feine et al., 1991). Our findings refute assumptions that female participants are less reliable due to hormonal fluctuations. The present study was not designed to test whether these sex differences in pain reliability are biological, learned, or reflective of the testing context (e.g. sex of the experimenter), although prior work suggests that neither experimenter gender nor type of rating scale is likely to influence sex differences in pain sensitivity or reliability (Feine et al., 1991). 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.

In contrast to previous QST reliability studies, our study included a wide range in the intervals between tests, which allowed us to test whether responses vary as a function of time between visits. We hypothesized that pain sensitivity measures would be more consistent within individuals when visits were closer in time. Interestingly, 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 (Moloney et al., 2012). We also tested whether pain sensitivity measures showed any consistent effects across participants as a function of visit number. We found that pain tolerance decreased across visits, and goodness-of-fit also decreased across time in the outpatient clinic, but not in the fMRI environment. 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 (Backonja et al., 2009), or the psychosocial context surrounding sensory testing, such as coherence with experimenter based on ethnicity or gender (Aslaksen et al., 2007; Losin et al., 2017).

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. As mentioned above, we focused on measures of supra-threshold pain, while other studies have focused primarily on the reliability of thermal detection 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 (LaMotte & Campbell, 1978), 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 health state evaluations (González-Fernández et al., 2014). 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 or other pain measures (e.g. pain biosignatures (Han et al., 2021)). We also found that linear models provided better fits than non-linear models in this rather large dataset, whereas previous work has suggested that the relationship between temperature and pain rating is nonlinear (Stevens, 1961; Sturgeon et al., 2015). 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. Finally, we evaluated reliability in data that were collected over the span of nearly five years, including different experimenters and testing environments. A 2020 study found that stability of experimenter is “extremely important” for interpretation of results in studies of test-retest reliability (W. Lin et al., 2020). 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.

Our study raises important outstanding questions that should be addressed in future work. First, we measured ASC task reliability in healthy volunteers, and need to formally evaluate whether this task or similar adaptive calibrations aree reliable in patient populations. 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 (e.g. 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 Table 1). Thermal pain thresholds have also been shown to have different levels of reliability on different areas of the body (Nothnagel et al., 2017). We did see slightly lower goodness-of-fit on the study that tested the calf relative to the studies that tested the forearm (see Table 1), 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.

In conclusion, our study examined the reliability of several measures of suprathreshold pain perception in a large sample of healthy volunteers who underwent an adaptive staircase heat pain calibration on multiple visits. Thermal pain threshold and tolerance were moderately reliable within and between individuals and remained relatively stable independent of sex, 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, individuals showed low reliability in the goodness-of-fit between temperature and pain. This suggests that goodness-of-fit is more sensitive to contextual factors that vary over visits, although individuals with strong associations showed high agreement over time. Our conclusions were consistent across multiple analytic approaches and whether we assumed linear or nonlinear associations between temperature and pain. 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 contextual pain modulation should continue to understand the contextual factors that contribute to variability.

Acknowledgements

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). 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. We have no conflicts of interest to report.

References

Arnegard, M. E., Whitten, L. A., Hunter, C., & Clayton, J. A. (2020). Sex as a Biological Variable: A 5-Year Progress Report and Call to Action. *Journal of Women's Health*, *29*(6), 858–864. https://doi.org/10.1089/jwh.2019.8247

Aslaksen, P. M., Myrbakk, I. N., Høifødt, R. S., & Flaten, M. A. (2007). The effect of experimenter gender on autonomic and subjective responses to pain stimuli. *Pain*, *129*(3), 260–268.

Atlas, L. Y., Bolger, N., Lindquist, M. A., & Wager, T. D. (2010). Brain Mediators of Predictive Cue Effects on Perceived Pain. *Journal of Neuroscience*, *30*(39), 12964–12977. https://doi.org/10.1523/JNEUROSCI.0057-10.2010

Atlas, L. Y., Lindquist, M. A., Bolger, N., & Wager, T. D. (2014). Brain mediators of the effects of noxious heat on pain. *Pain*, *155*(8), 1632–1648. https://doi.org/10.1016/j.pain.2014.05.015

Atlas, L. Y., Whittington, R. A., Lindquist, M. A., Wielgosz, J., Sonty, N., & Wager, T. D. (2012). Dissociable Influences of Opiates and Expectations on Pain. *Journal of Neuroscience*, *32*(23), 8053–8064. https://doi.org/10.1523/JNEUROSCI.0383-12.2012

Backonja, M.-M., Walk, D., Edwards, R. R., Sehgal, N., Moeller-Bertram, T., Wasan, A., Irving, G., Argoff, C., & Wallace, M. (2009). Quantitative Sensory Testing in Measurement of Neuropathic Pain Phenomena and Other Sensory Abnormalities. *The Clinical Journal of Pain*, *25*(7), 641–647. https://doi.org/10.1097/AJP.0b013e3181a68c7e

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations*, *9*, 205979911667287. https://doi.org/10.1177/2059799116672875

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, *8*, 307–310.

Bland, J. M., & Altman, D. G. (1996). Statistics Notes: Measurement error proportional to the mean. *BMJ*, *313*(7049), 106–106. https://doi.org/10.1136/bmj.313.7049.106

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*, 135–160.

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: What is it, and how is it measured? *Physiotherapy*, *86*(2), 94–99.

Caldwell, A. (2021). *SimplyAgree: Flexible and Robust Agreement and Reliability Analyses.* (R package version 0.0.2) [Computer software].

Canty, A., & Ripley, B. (2020). *boot: Bootstrap R (S-Plus) Functions* (R package version 1.3-25) [Computer software].

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Clayton, J. A. (2018). Applying the new SABV (sex as a biological variable) policy to research and clinical care. *Physiology & Behavior*, *187*, 2–5. https://doi.org/10.1016/j.physbeh.2017.08.012

Clayton, J. A., & Collins, F. S. (2014). Policy: NIH to balance sex in cell and animal studies. *Nature*, *509*(7500), 282–283. https://doi.org/10.1038/509282a

Davis, K. D., Aghaeepour, N., Ahn, A. H., Angst, M. S., Borsook, D., Brenton, A., Burczynski, M. E., Crean, C., Edwards, R., Gaudilliere, B., Hergenroeder, G. W., Iadarola, M. J., Iyengar, S., Jiang, Y., Kong, J.-T., Mackey, S., Saab, C. Y., Sang, C. N., Scholz, J., … Pelleymounter, M. A. (2020). Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: Challenges and opportunities. *Nature Reviews Neurology*, *16*(7), 381–400. https://doi.org/10.1038/s41582-020-0362-2

Davis, K. D., & Cheng, J. C. (2019). Differentiating trait pain from state pain: A window into brain mechanisms underlying how we experience and cope with pain. *PAIN Reports*, *4*(4), e735. https://doi.org/10.1097/PR9.0000000000000735

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (Issue 1). Cambridge university press.

Dildine, T. C., Necka, E. A., & Atlas, L. Y. (2020). Confidence in subjective pain is predicted by reaction time during decision making. *Scientific Reports*, *10*(1), 21373. https://doi.org/10.1038/s41598-020-77864-8

Feine, J. S., Bushnell, C. M., Miron, D., & Duncan, G. H. (1991). Sex differences in the perception of noxious heat stimuli. *Pain*, *44*(3), 255–262. https://doi.org/10.1016/0304-3959(91)90094-E

Feldhaus, M. H., Horing, B., Sprenger, C., & Büchel, C. (2021). Association of nocebo hyperalgesia and basic somatosensory characteristics in a large cohort. *Scientific Reports*, *11*(1), 762. https://doi.org/10.1038/s41598-020-80386-y

Felix, E. R., & Widerström-Noga, E. G. (2009). Reliability and validity of quantitative sensory testing in persons with spinal cord injury and neuropathic pain. *Journal of Rehabilitation Research & Development*, *46*(1).

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). Package 'irr.' *Various Coefficients of Interrater Reliability and Agreement*, *22*.

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. (2011). 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*(3), 548–556. https://doi.org/10.1016/j.pain.2010.11.013

González-Fernández, M., Ghosh, N., Ellison, T., McLeod, J. C., Pelletier, C. A., & Williams, K. (2014). Moving Beyond the Limitations of the Visual Analog Scale for Measuring Pain: Novel Use of the General Labeled Magnitude Scale in a Clinical Setting. *American Journal of Physical Medicine & Rehabilitation*, *93*(1), 75–81. https://doi.org/10.1097/PHM.0b013e31829e76f7

Grahl, A., Onat, S., & Büchel, C. (2018). The periaqueductal gray and Bayesian integration in placebo analgesia. *ELife*, *7*, e32930. https://doi.org/10.7554/eLife.32930

Green, B., & Cruz, A. (1998). “Warmth-insensitive fields”: Evidence of sparse and irregular innervation of human skin by the warmth sense. *Somatosensory & Motor Research*, *15*(4), 269–275. https://doi.org/10.1080/08990229870682

Han, X., Ashar, Y. K., Kragel, P., Petre, B., Schelkun, V., Atlas, L., Chang, L. J., Jepma, M., Koban, L., Losin, E. R., Roy, M., Woo, C.-W., & Wager, T. D. (2021). *Effect sizes and test-retest reliability of the fMRI-based Neurologic Pain Signature*. https://doi.org/10.1101/2021.05.29.445964

Hansson, P., Backonja, M., & Bouhassira, D. (2007). Usefulness and limitations of quantitative sensory testing: Clinical and research application in neuropathic pain states. *Pain*, *129*(3), 256–259. https://doi.org/10.1016/j.pain.2007.03.030

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. https://doi.org/10.1016/j.jcm.2016.02.012

Kragel, P. A., Han, X., Kraynak, T. E., Gianaros, P. J., & Wager, T. D. (2021). Functional MRI Can Be Highly Reliable, but It Depends on What You Measure: A Commentary on Elliott et al. (2020). *Psychological Science*, *32*(4), 622–626. https://doi.org/10.1177/0956797621989730

LaMotte, R. H., & Campbell, J. N. (1978). Comparison of responses of warm and nociceptive C-fiber afferents in monkey with human judgments of thermal pain. *Journal of Neurophysiology*, *41*(2), 509–528. https://doi.org/10.1152/jn.1978.41.2.509

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Lin, L. I.-K. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, *45*(1), 255. https://doi.org/10.2307/2532051

Lin, W., Zhou, F., Yu, L., Wan, L., Yuan, H., Wang, K., & Svensson, P. (2020). Quantitative sensory testing of periauricular skin in healthy adults. *Scientific Reports*, *10*(1), 3728. https://doi.org/10.1038/s41598-020-60724-w

Losin, E. A. R., Anderson, S. R., & Wager, T. D. (2017). Feelings of Clinician-Patient Similarity and Trust Influence Pain: Evidence From Simulated Clinical Interactions. *The Journal of Pain*, *18*(7), 787–799. https://doi.org/10.1016/j.jpain.2017.02.428

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60).

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Mischkowski, D., Palacios-Barrios, E. E., Banker, L., Dildine, T. C., & Atlas, L. Y. (2019). Pain or nociception? Subjective experience mediates the effects of acute noxious heat on autonomic responses - corrected and republished. *Pain*, *160*(6), 1469–1481. https://doi.org/10.1097/j.pain.0000000000001573

Mischkowski, D., Stavish, C. M., Palacios-Barrios, E. E., Banker, L. A., Dildine, T. C., & Atlas, L. Y. (2021). Dispositional Mindfulness and Acute Heat Pain: Comparing Stimulus-Evoked Pain With Summary Pain Assessment. *Psychosomatic Medicine*, *83*(6), 539–548. https://doi.org/10.1097/PSY.0000000000000911

Mogil, J. S., & Chanda, M. L. (2005). The case for the inclusion of female subjects in basic science studies of pain. *Pain*, *117*(1), 1–5. https://doi.org/10.1016/j.pain.2005.06.020

Moloney, N. A., Hall, T. M., & Doody, C. M. (2012). Reliability of thermal quantitative sensory testing: A systematic review. *The Journal of Rehabilitation Research and Development*, *49*(2), 191. https://doi.org/10.1682/JRRD.2011.03.0044

Nothnagel, H., Puta, C., Lehmann, T., Baumbach, P., Menard, M. B., Gabriel, B., Gabriel, H. H., Weiss, T., & Musial, F. (2017). How stable are quantitative sensory testing measurements over time? Report on 10-week reliability and agreement of results in healthy volunteers. *Journal of Pain Research*. https://doi.org/10.2147/JPR.S137391

Pigg, M., Baad-Hansen, L., Svensson, P., Drangsholt, M., & List, T. (2010). Reliability of intraoral quantitative sensory testing (QST). *Pain*, *148*(2), 220–226.

Pleil, J. D., Wallace, M. A. G., Stiegel, M. A., & Funk, W. E. (2018). Human biomarker interpretation: The importance of intra-class correlation coefficients (ICC) and their calculations based on mixed models, ANOVA, and variance estimates. *Journal of Toxicology and Environmental Health, Part B*, *21*(3), 161–180. https://doi.org/10.1080/10937404.2018.1490128

Prendergast, B. J., Onishi, K. G., & Zucker, I. (2014). Female mice liberated for inclusion in neuroscience and biomedical research. *Neuroscience & Biobehavioral Reviews*, *40*, 1–5. https://doi.org/10.1016/j.neubiorev.2014.01.001

Qin, S., Nelson, L., McLeod, L., Eremenco, S., & Coons, S. J. (2019). Assessing test–retest reliability of patient-reported outcome measures using intraclass correlation coefficients:

Recommendations for selecting and documenting the analytical formula. *Quality of Life Research*, *28*(4), 1029–1033. https://doi.org/10.1007/s11136-018-2076-0

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rolke, R., Baron, R., Maier, C., Tölle, T. R., Treede, - D. R., Beyer, A., Binder, A., Birbaumer, N., Birklein, F., Bötefür, I. C., Braune, S., Flor, H., Huge, V., Klug, R., Landwehrmeyer, G. B., Magerl, W., Maihöfner, C., Rolko, C., Schaub, C., … Wasserka, B. (2006). Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): Standardized protocol and reference values. *Pain*, *123*(3), 231–243. https://doi.org/10.1016/j.pain.2006.01.041

Searle, S. R. (1971). *Linear Models*. Wiley.

Shansky, R. M. (2018). Sex differences in behavioral strategies: Avoiding interpretational pitfalls. *Current Opinion in Neurobiology*, *49*, 95–98. https://doi.org/10.1016/j.conb.2018.01.007

Shansky, R. M. (2019). Are hormones a “female problem” for animal research? *Science*, *364*(6443), 825–826. https://doi.org/10.1126/science.aaw7570

Shansky, R. M., & Murphy, A. Z. (2021). Considering sex as a biological variable will require a global shift in science culture. *Nature Neuroscience*, *24*(4), 457–464. https://doi.org/10.1038/s41593-021-00806-8

Shih, Y.-W., Tsai, H.-Y., Lin, F.-S., Lin, Y.-H., Chiang, C.-Y., Lu, Z.-L., & Tseng, M.-T. (2019). Effects of Positive and Negative Expectations on Human Pain Perception Engage Separate But Interrelated and Dependently Regulated Cerebral Mechanisms. *The Journal of Neuroscience*, *39*(7), 1261–1274. https://doi.org/10.1523/JNEUROSCI.2154-18.2018

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181. https://doi.org/10.1037/h0046162

Stevens, S. S. (1961). To Honor Fechner and Repeal His Law: A power function, not a log function, describes the operating characteristic of a sensory system. *Science*, *133*(3446), 80–86. https://doi.org/10.1126/science.133.3446.80

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2017). RptR: Repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *8*(11), 1639–1644.

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2019, March 6). *An introduction to repeatability estimation with rptR*. https://cran.r-project.org/web/packages/rptR/vignettes/rptR.html

Sturgeon, J. A., Tieu, M. M., Jastrzab, L. E., McCue, R., Gandhi, V., & Mackey, S. C. (2015). Nonlinear Effects of Noxious Thermal Stimulation and Working Memory Demands on Subjective Pain Perception. *Pain Medicine*, *16*(7), 1301–1310. https://doi.org/10.1111/pme.12774

Wasner, G. L., & Brock, J. A. (2008). Determinants of thermal pain thresholds in normal subjects. *Clinical Neurophysiology*, *119*(10), 2389–2395.

Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for estimating repeatability. *Methods in Ecology and Evolution*, *3*(1), 129–137.

Yarnitsky, D., Sprecher, E., Zaslansky, R., & Hemli, J. A. (1995). Heat pain thresholds: Normative data and repeatability. *Pain*, *60*(3), 329–332.

Zhang, S., Yoshida, W., Mano, H., Yanagisawa, T., Mancini, F., Shibata, K., Kawato, M., & Seymour, B. (2020). Pain Control by Co-adaptive Learning in a Brain-Machine Interface. *Current Biology*, *30*(20), 3935-3944.e7. https://doi.org/10.1016/j.cub.2020.07.066

Tables

Table 1. Participation and pain sensitivity by visit type.[a]

Visit type	N	Body site	Environment	Non-pain stimuli?	Threshold	Tolerance	Goodness of fit (r^2), all trials	Goodness of fit (r^2), outliers removed
Screening / ASC only	318	Left arm	Clinic testing room	N	42.00 (3.17)	48.66 (3.92)	0.67 (0.18)	0.79 (0.20)
ASC followed by cued attention task	23	Left leg	Clinic testing room	N	42.61 (1.82)	47.33 (1.99)	0.53 (0.23)	0.63 (0.25)
ASC followed by learning task	12	Left arm	Clinic testing room	N	41.13 (3.35)	47.79 (2.53)	0.70 (0.09)	0.82 (0.11)
ASC followed by expectancy fMRI task	47	Left arm	fMRI suite	N	42.18 (2.57)	47.21 (1.96)	0.65 (0.19)	0.77 (0.18)
ASC for heat and tastants	30	Left arm	Clinic testing room	Sugar, Salt, neutral taste	42.80 (2.63)	47.87 (2.05)	0.63 (0.19)	0.75 (0.21)
ASC for heat and tastants followed by expectancy task	73	Left arm	Clinic testing room	Sugar, Salt, neutral taste	42.40 (1.99)	48.19 (1.70)	0.73 (0.12)	0.75 (0.13)
ASC followed by expectancy fMRI	21	Left arm	fMRI suite	N	42.82 (1.35)	48.60 (1.58)	0.75 (0.11)	0.77 (0.11)
ASC followed by placebo experiment	17	Left arm	Clinic testing room	N	41.59 (1.65)	47.29 (1.77)	0.72 (0.15)	0.84 (0.14)
ASC followed by placebo fMRI study	65	Left arm	fMRI suite	N	42.95 (2.32)	47.98 (1.70)	0.71 (0.14)	0.81 (0.15)

[a] This table reports threshold, tolerance, and goodness-of-fit as a function of visit type, based on an adaptive staircase calibration (ASC) administered on each visit. The column labeled “N” reports the number of ASC visits per study at each visit, and the four rightmost columns report mean and standard deviation for each measure. All stimuli were 8s duration with the exception of 12 Screening Visit participants whose stimuli were 10s in duration.

Table 2. Participation and pain sensitivity by visit number.[b]

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), outliers removed
1	All participants		342	*M* = 41.92, *SD* = 3.16	*M* = 48.59, *SD* = 3.81	*M* = 0.67, *SD* = 0.18	*M* = 0.78, *SD* = 0.19
		Participants with multiple visits	171	*M* = 41.83, *SD* = 2.42	*M* = 48.23, *SD* = 1.77	*M* = 0.72, *SD* = 0.13	*M* = 0.85, *SD* = 0.11
		Participants with one visit	171	*M* = 42.02, *SD* = 3.73	*M* = 48.95, *SD* = 5.07	*M* = 0.63, *SD* = 0.21	*M* = 0.72, *SD* = 0.22
2			171	*M* = 42.40, *SD* = 2.25	*M* = 47.66, *SD* = 1.84	*M* = 0.68, *SD* = 0.18	*M* = 0.76, *SD* = 0.18
3			63	*M* = 42.81, *SD* = 2.21	*M* = 48.17, *SD* = 2.09	*M* = 0.67, *SD* = 0.18	*M* = 0.77, *SD* = 0.16
4			22	*M* = 43.57, *SD* = 1.47	*M* = 48.24, *SD* = 1.62	*M* = 0.70, *SD* = 0.18	*M* = 0.76, *SD* = 0.21
5			6	*M* = 43.16, *SD* = 1.76	*M* = 47.88, *SD* = 1.66	*M* = 0.67, *SD* = 0.09	*M* = 0.77, *SD* = 0.14

[b] 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 3.[c]

	Threshold		Tolerance		Goodness-of-fit (including outliers)	
	ICC	CI	ICC	CI	ICC	CI
Male	0.592	[0.46, 0.71]	0.555	[0.41, 0.68]	0.064	[-.09, 0.24]
Female	0.679	[0.58, 0.77]	0.708	[0.61, 0.79]	0.318	[0.17, 0.46]

[c] 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 (Wolak et al., 2012), with confidence intervals based on Searle (Searle, 1971).