**Effect size and test-retest reliability of the fMRI-based Neurologic Pain Signature**

Xiaochun Han1

Yoni K. Ashar2

Philip Kragel3

Bogdan Petre1

Victoria Schelkun4

Lauren Atlas5-7

Luke Chang1

Marieke Jepma8

Leonie Koban9

Elizabeth Reynolds Losin10

Mathieu Roy11

Choong-Wan Woo12

Tor D. Wager1

1Dartmouth College, Hanover, NH

2Weill Cornell Medical College, New York, NY

3Emory University, Atlanta, GA

4Columbia University, New York, NY

5National Center for Complementary and Integrative Health, National Institutes of Health, Bethesda, MD

6National Institute of Mental Health, National Institutes of Health, Bethesda, MD 7National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD

8University of Amsterdam, Amsterdam

9INSEAD Fontainebleau & ICM paris, Paris

10Department of Psychology, University of Miami, Miami FL

11Department of Psychology, McGill University, Montreal, Quebec

12Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, Gyeonggi-do

Running Head: NPS MEASUREMENT PROPERTIES

Please address correspondence to:

Tor D. Wager Professor

Department of Psychological and Brain Sciences

Dartmouth College

3 Maynard St.

Hanover, NH 03784

Email: Tor.D.Wager@Dartmouth.edu

Telephone: (303) 895-8739

SUMMARY FOR THIS DRAFT

4 figures; 6 supp. 2 Tables

Character count: 42825 characters with spaces

Figures: 7911 characters with spaces.

**Abstract**

Identifying biomarkers with large effect sizes and high test-retest reliability is a growing priority for fMRI research. We examined a well-established multivariate measure that tracks pain induced by nociceptive input, the Neurologic Pain Signature (NPS). In N = 294 participants across eight studies, meta-analyses of NPS responses showed a very-large effect size in predicting within-person single-trial pain reports (d = 1.45) and a medium effect size in predicting individual differences in pain reports (d = 0.49). The NPS showed excellent short-term (within-day) test-retest reliability (ICC = 0.84). Reliability scaled with the number of trials within-person, with ≥60 trials required for excellent test-retest reliability. Reliability was comparable in two additional studies across 5-day (N = 27) and 1-month (N = 40) test-retest intervals. The reliability of both pain reports and NPS expression exceeded the correlation between the two measures, indicating different sources of between-person variances and promise for multimodal, combined measures.

**Author Note**

This project was supported by grants R01MH076136 (T.D.W.), R01DA046064, R01EB026549, and R01DA035484. Code for all analyses and figures is available at: https://github.com/XiaochunHan/NPS\_measurement\_properties

Effect size and test-retest reliability of the fMRI-based Neurologic Pain Signature

Understanding individual differences in brain activity and their links with behavior is a primary focus of fMRI research. Thousands of studies have characterized correlations between MRI measures and individual differences in external variables, including cognitive performance [(Lund et al. 2020; Zhao et al. 2021)](https://paperpile.com/c/tXAZY0/XgU0+3WfH); mental health and neurological disorders [(Swartz et al. 2015; Gabrieli et al. 2015; Woo et al. 2017; Drysdale et al. 2017)](https://paperpile.com/c/tXAZY0/0taC+6aMb+XDxm+wj1A); age and development stage [(Cole and Franke 2017; Marek et al. 2019)](https://paperpile.com/c/tXAZY0/dlRH+NzaQ); personality and interpersonal characteristics  [(DeYoung et al. 2010; Maglanoc et al. 2020)](https://paperpile.com/c/tXAZY0/PTcg+nV7v); and genetics [(Reus et al. 2017; Neilson et al. 2019)](https://paperpile.com/c/tXAZY0/bdYh+J8UE). One goal of these studies is to establish structure-function associations, and make inferences about the brain bases of individual differences. A distinct, but related, goal is to use brain measures to develop biomarkers that can contribute to measuring external constructs (e.g., pain, risk for mental illness, etc.) and inform diagnosis and treatment [(FDA-NIH Biomarker Working Group, 2016)](https://paperpile.com/c/mO5B1U/4Xnln). For example, research on neural mechanisms of physiological pain help to understand which brain areas involved in constructing different kinds of pain experience in different populations; and to develop biomarkers that can subtype pain based on pathophysiology, predicting risk for future pain, and more, leading to new ways of understanding, diagnosing, and treating pain.

Both establishing structure-function associations and developing biomarkers require brain measures with good measurement properties, including large effect sizes in predicting external variables (e.g., behavior) and high reliability. Effect sizes in predicting external variables are calculated at both within-person and between-person levels when repeated measures are collected within each person. Two levels of predictions might be inconsistent, known as Simpson’s paradox, due to different sources of variance at different levels [(Kievit et al. 2013; Bakdash and Marusich 2017)](https://paperpile.com/c/tXAZY0/12rG+LSJj). For example, at the between-person level, people with more alcohol intake have higher intelligence; while at the within-person level, people with more alcohol intake have lower intelligence. To test effect sizes at both within-person and between-person levels prevents incorrect interpretations of the predictions and results in a deeper understanding of the brain measures. Test-retest reliability is one type of reliability indexes, usually measured with intraclass correlation coefficient (ICC, [Shrout and Fleiss 1979](https://paperpile.com/c/tXAZY0/1NUQ)), that assesses temporal stability under repeated tests. Both effect sizes and test-retest reliability rely on low random error in the measurement. Test-retest reliability also relies on high inter-individual variability, indicating differentiable measures across subjects [(Barnhart, Haber, and Lin 2007)](https://paperpile.com/c/mO5B1U/Rv2Wf).

Historically, the goal of maximizing effect size and reliability has been implicit in most fMRI studies, and these properties have seldom been assessed directly. As translational goals accelerate and sample sizes increase, measurement properties of fMRI are increasingly a focus of attention [(Button et al. 2013; Poldrack et al. 2017; Nichols et al. 2017; Bennett and Miller 2010; Kraemer 2014; Zuo and Xing 2014; Dubois and Adolphs 2016; Xu et al. 2016; O'Connor et al. 2017; Hedge et al. 2018; Herting et al. 2018; Elliott et al. 2019; Noble et al. 2019; Zuo et al. 2019; Elliott et al. 2020)](https://paperpile.com/c/mO5B1U/O7Qe+K0Ve+YeBm+UDwQ+VKSp+mEW2+2BgL+Lc97+ORAB+soRh+DhDa+wDkT+5khp+rPLR+8Elx).Studies of traditional univariate brain measures provide a pessimistic picture of task fMRI’s measurement properties. Effect sizes of univariate brain measures in local brain regions have often been limited to moderate effect sizes (i.e., Cohen’s d values centered on approximately d = 0.5; [Poldrack, Baker & Durnez, 2017](https://paperpile.com/c/mO5B1U/K0Ve)). The reliability of univariate brain measures in many studies with small samples varied substantially [(Manuck et al. 2007; Plichta et al. 2012; Nord et al. 2017; Letzen et al. 2016)](https://paperpile.com/c/tXAZY0/e2Gk+JsQg+3kL7+VHyr). A recent meta-analysis of fMRI literature across diverse tasks generally demonstrated low reliability (ICCs < 0.4) of the average activation level of single brain regions of interest (ROI), which did not decrease with longer test-retest interval [(Elliott et al. 2020)](https://paperpile.com/c/mO5B1U/8Elx). In resting-state fMRI studies, low test-retest reliability (ICCs < 0.3) was indicated in the individual edge-level connectivity [(Pannunzi et al., 2017; Noble, Scheinost & Constable, 2019)](https://paperpile.com/c/mO5B1U/xMxrX+5khp).

An important trend in the field is the development of *a priori* multivariate brain measures that can be used as biomarkers, also called ‘neuromarkers’ or ‘signatures’ [(Orrù et al. 2012; Doyle et al. 2015; Haynes 2015; Gabrieli et al. 2015; Arbabshirani et al. 2017; Woo et al. 2017; Kragel et al. 2018)](https://paperpile.com/c/mO5B1U/YcywJ+gRhIe+PSHYk+sFiQ+mHb6+gxX3D+IrEW1). Such models consist of patterns of brain activity, connectivity, and other derived features (e.g., graph theoretic measures) within and across brain regions, which can be applied prospectively to new samples and/or participants. Because they are pre-specified models that are applied to new samples without re-fitting, neuromarkers provide an opportunity to systematically evaluate measurement properties across different samples and contexts. Multivariate brain signatures can yield measures with much larger effect sizes (Cohen’s d > 2; [Wager et al., 2013; Zunhammer et al., 2018; Chang et al., 2015; Krishnan et al., 2016; Geuter et al., 2020](https://paperpile.com/c/mO5B1U/EIYNO+i3j8M+zGR3c+xKdrz+4nfsj)). They also show enhanced test-retest reliability for both task-evoked (ICCs > 0.7; [Woo & Wager, 2016; Kragel et al., 2020](https://paperpile.com/c/mO5B1U/DkZcQ+R9WxY)) and resting-state (ICCs > 0.6; [Zuo and Xing 2014; Yoo et al. 2019](https://paperpile.com/c/mO5B1U/mEW2+O7M4q)) fMRI measures in some studies. However, this has rarely been done across diverse samples and scanners, particularly with respect to systematic evaluation of effect sizes for within-person and between-person prediction of external variables and test-retest reliability. That is the goal here.

In the current study, we evaluated a well-established multivariate brain-based model in the pain domain, i.e., the Neurologic Pain Signature (NPS; [Wager et al., 2013](https://paperpile.com/c/mO5B1U/EIYNO)). NPS consists of interpretable and stable patterns across brain regions known to show increased activity in pain-related studies. These regions included the thalamus, the posterior and anterior insulae, the secondary somatosensory cortex, the anterior cingulate cortex, the periaqueductal gray matter, and other regions (see **Figure 1(A)**;[Yarkoni et al., 2011](https://paperpile.com/c/mO5B1U/tKr4e)). The NPS predicts subjective pain intensity in response to noxious thermal [(Wager et al., 2013)](https://paperpile.com/c/mO5B1U/EIYNO), mechanical [(Krishnan et al., 2016)](https://paperpile.com/c/mO5B1U/xKdrz), electrical [(Krishnan et al., 2016; Ma et al., 2016)](https://paperpile.com/c/mO5B1U/xKdrz+wLkF8), and visceral stimuli [(Van Oudenhove et al. 2020)](https://paperpile.com/c/mO5B1U/g9Kp). In addition, it does not respond to non-noxious warm stimuli [(Wager et al., 2013)](https://paperpile.com/c/mO5B1U/EIYNO), threat cues [(Wager et al., 2013; Krishnan et al., 2016; Ma et al., 2016)](https://paperpile.com/c/mO5B1U/EIYNO+xKdrz+wLkF8), social rejection-related stimuli [(Wager et al., 2013)](https://paperpile.com/c/mO5B1U/EIYNO), vicarious pain [(Krishnan et al., 2016)](https://paperpile.com/c/mO5B1U/xKdrz), or aversive images [(Chang et al., 2015)](https://paperpile.com/c/mO5B1U/zGR3c). The signature does not measure a disorder but rather provides a neuromarker of a basic mental process with negative sensory and affective components, which can serve as an intermediate phenotype, which can be altered in various disorders. For example, enhanced NPS responses, combined with another brain signature related to non-painful sensory processing, discriminated fibromyalgia from pain-free controls with 93% accuracy [(López-Solà et al., 2017)](https://paperpile.com/c/mO5B1U/ZA7dX). However, its effect sizes have been mainly assessed on within-person correlations with pain [(Lindquist et al. 2017)](https://paperpile.com/c/mO5B1U/UCbF), and reliability has only been assessed in a preliminary fashion [(Woo and Wager 2016; Kragel et al. 2020)](https://paperpile.com/c/mO5B1U/DkZcQ+R9WxY). Properties that influence its test-retest reliability (e.g., amount of data collected per person) has not been systematically examined in detail across studies. Examining these properties could both help understand the NPS as a test case and reveal principles underlying the sources of error and reliability of task fMRI more broadly.

We tested four types of effect size for both NPS and local brain regions of interest by analyzing painful stimulus-evoked fMRI and pain reports across 10 studies (total N = 441), none of which were used to train the NPS model. The main analyses were conducted on single trial-level data from a multi-study dataset across 8 studies (N = 294).  With this dataset, we also tested short-term (i.e., within one-day) test-retest reliability. We also tested several factors that might affect the short-term test-retest reliability [(Bennett & Miller 2010, 2013)](https://paperpile.com/c/mO5B1U/UDwQ+zgbd), including the number of trials used, the noxious stimulus intensity, and whether the NPS was applied to pain-versus-rest or a contrast between high and low painful stimulus intensity. Studies 9 and 10 evaluated test-retest reliability across 5-days (N = 27) and one-month (N = 120) intervals. We compared the short-term and long-term test-retest reliability controlling for the number of trials averaged when calculating the NPS response.

**Methods**

**Datasets description**

We analyzed three datasets: (1) a single-trial dataset on healthy subjects during heat pain tasks with behavioral and fMRI data collected within one day (i.e., one session). The single-trial dataset included 15,904 single-trial images of fMRI activity associated with multiple levels of noxious heat and pain ratings across 294 participants from 8 studies. (2) a study on healthy subjects during heat pain tasks with behavioral and fMRI data collected across three sessions with five-days interval on average (N = 27). (3) a study on chronic pain subjects during pressure pain tasks with behavioral and fMRI data collected across two sessions with one-month interval on average (N = 120). In all studies, participants received a series of painful stimuli and rated their individually experienced pain following each stimulus. Each study also comprises a specific psychological manipulation, such as cue-induced expectation and placebo treatment. Descriptive data on age, sex, and other study sample features are given in **Table 1**. The number of trials, stimulation sites, rating scales, and stimulus intensities and durations varied across studies but were comparable; these variables are summarized in **Table 2**. In the studies included, we examined the test-retest reliability of the NPS and pain ratings irrespective of diverse study-specific features and manipulations, which facilitated our conclusion's generalizability.

Data from the single-trial dataset have been used in previous publications (see **Table 1**). However, the analyses and findings reported here are novel, and no data for NPS development is included in the current study to avoid double-dipping. Data from the other two studies have not been published yet. All participants were recruited from New York City and Boulder/Denver Metro Areas. The institutional review board of Columbia University and the University of Colorado Boulder approved all the studies, and all participants provided written informed consent. Participants' preliminary eligibility was determined through an online questionnaire, a pain safety screening form, and an MRI safety screening form. Participants with psychiatric, physiological, or pain disorders, neurological conditions, and MRI contraindications were excluded before enrollment. No participants were excluded from the study after screening other than individuals who, upon screening, provided different responses that made them ineligible (e.g., developing a physiological disorder).

**Materials and Procedures**

**Thermal and pressure stimulation**

We delivered thermal stimulation to multiple skin sites using a TSA-II Neurosensory Analyzer (Medoc Ltd., Chapel Hill, NC) with a 16 mm Peltier thermode endplate (Study 8 (scebl): 32 mm). One study delivered pressure rather than thermal stimulation, using a custom-built pneumatic device pushing a piston into the left thumbnail. At the end of every trial, participants rated pain intensity on a visual analog scale or a labeled magnitude scale [(Bartoshuk et al., 2004)](https://paperpile.com/c/mO5B1U/jJXnP). Thermal stimulation parameters varied across studies, with stimulation temperatures ranging from 44.3 °C to 50 °C and stimulation durations ranging from 1.85 to 12.5 seconds. Most studies applied thermal stimulation to the glabrous skin of the left forearm; study 2 (bmrk4) additionally to the dorsum of the left foot, study 6 (ie2) and study 8 (scebl) applied the stimulation to the lower leg. See **Table 2** for stimulation location, intensity levels, duration, number of trials per subject, and other cognitive manipulations.

**fMRI Analysis**

**Preprocessing of the single-trial dataset**

Structural T1-weighted images were coregistered to each subject's mean functional image using the iterative mutual information-based algorithm implemented in SPM [(Ashburner & Friston, 2005)](https://paperpile.com/c/mO5B1U/hLY9u). They were then normalized to MNI space using SPM. Following SPM normalization, study 4 (exp) included an additional step of normalization to the group mean using a genetic algorithm-based normalization [(Atlas et al., 2010, 2014; Wager & Nichols, 2003)](https://paperpile.com/c/mO5B1U/4jqzz+leL5Q+LTtbt). In each functional run, we removed initial volumes to allow for image intensity stabilization. We also identified image-intensity outliers (i.e., 'spikes') by computing the mean and standard deviations (SD, across voxels) of intensity values for each image for all slices to remove intermittent gradient and severe motion-related artifacts present to some degree in all fMRI data. We first computed both the mean and the SD of intensity values across each slice for each image to identify outliers. Mahalanobis distances for the matrix of (concatenated) slice-wise mean and standard deviation values by functional volumes (over time) were computed. Any values with a significant χ2 value (corrected for multiple comparisons based on the more stringent of either false discovery rate or Bonferroni methods) were considered outliers. In practice, less than 1% of the images were deemed outliers. The outputs of this procedure were later included as nuisance covariates in the first level models. Next, functional images were corrected for differences in each slice's acquisition timing and were motion-corrected (realigned) using SPM. The functional images were warped to SPM's normative atlas (warping parameters estimated from coregistered, high-resolution structural images), interpolated to 2 × 2 × 2 mm3 voxels, and smoothed with an 8 mm FWHM Gaussian kernel. This smoothing level has been shown to improve inter-subject functional alignment while retaining sensitivity to mesoscopic activity patterns consistent across individuals [(Shmuel et al., 2010)](https://paperpile.com/c/mO5B1U/a77ha).

**Preprocessing with *fMRIPrep***

The preprocessing of study 10 (rtnf) and study 11 (olp4cbp) were conducted using *fMRIPrep* 1.2.4 ([Esteban et al., 2019](https://paperpile.com/c/mO5B1U/Qjx4Z); Esteban, Blair, et al., 2018; RRID:SCR\_016216), which is based on *Nipype* 1.1.6 ([Gorgolewski et al., 2011](https://paperpile.com/c/mO5B1U/FbKUV); Gorgolewski et al., 2018; RRID:SCR\_002502). T1-weighted images were corrected for intensity non-uniformity (INU) using *N4BiasFieldCorrection* ([Tustison et al., 2010](https://paperpile.com/c/mO5B1U/XE6EL); ANTs 2.2.0). A T1w-reference map was computed after registration of 3 T1w images (after INU-correction) using *mri\_robust\_template* (FreeSurfer 6.0.1; [Reuter, Rosas & Fischl, 2010](https://paperpile.com/c/mO5B1U/xiYrB)). The T1w-reference was then skull-stripped using *antsBrainExtraction.sh* (ANTs 2.2.0), using OASIS as target template. Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c ([Fonov et al., 2009](https://paperpile.com/c/mO5B1U/zPB2H); RRID:SCR\_008796) was performed through nonlinear registration with *antsRegistration* (ANTs 2.2.0, RRID:SCR\_004757; [Avants et al., 2008](https://paperpile.com/c/mO5B1U/JzNP8)), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using *fast* (FSL 5.0.9, RRID:SCR\_002823; [Zhang, Brady & Smith, 2001)](https://paperpile.com/c/mO5B1U/xXWDh).

For the functional images, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. The BOLD reference was then co-registered to the T1w reference using *flirt* (FSL 5.0.9; [Jenkinson & Smith, 2001)](https://paperpile.com/c/mO5B1U/41ArI) with the boundary-based registration [(Greve & Fischl, 2009)](https://paperpile.com/c/mO5B1U/PJqdc) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using *mcflirt* (FSL 5.0.9; [Jenkinson et al., 2002)](https://paperpile.com/c/mO5B1U/HrfvI). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled to *MNI152NLin2009cAsym* standard space, generating a *preprocessed BOLD* run in *MNI152NLin2009cAsym* space. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*; [Behzadi et al., 2007](https://paperpile.com/c/mO5B1U/cuxMN)). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). Six tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, six components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and template spaces). Gridded (volumetric) resamplings were performed using *antsApplyTransforms* (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels [(Lanczos, 1964)](https://paperpile.com/c/mO5B1U/wluL4). Non-gridded (surface) resamplings were performed using *mri\_vol2surf* (FreeSurfer).

**Single trial analyses**

For study 1 to 10, a single trial, or "single-epoch", design and analysis approach was employed to model the data. Quantification of single-trial response magnitudes was done by constructing a GLM design matrix with separate regressors for each trial, as in the "beta series" approach [(Mumford et al., 2012; Rissman, Gazzaley & D'Esposito, 2004)](https://paperpile.com/c/mO5B1U/3TN79+jWZCs). First, boxcar regressors, convolved with the canonical hemodynamic response function (HRF), were constructed to model cue, pain, and rating periods in each study. Then, we included a regressor for each trial, as well as several types of nuisance covariates. Because each trial consisted of relatively few volumes, trial estimates could be strongly affected by acquisition artifacts that occur during that trial (e.g., sudden motion, scanner pulse artifacts). Therefore, trial-by-trial variance inflation factors (VIFs; a measure of design-induced uncertainty due, in this case, to collinearity with nuisance regressors) were calculated, and any trials with VIFs that exceeded 2.5 were excluded from the analyses. For Study 8 (nsf), we also excluded global outliers (trials that exceeded three SDs above the mean) and employed a principal component-based denoising-step during preprocessing to minimize artifacts. This approach generated single-trial estimates that reflect the amplitude of the fitted HRF on each trial and refer to the magnitude of anticipatory and pain-period activity for each trial in each voxel. Single-trial analysis for Study 2 (bmrk4) and Study 4 (exp) were based on fitting a set of three basis functions, rather than the standard HRF used in the other studies. This flexible strategy allowed the shape of the modeled hemodynamic response function (HRF) to vary across trials and voxels. This procedure differed from that used in other studies because (a) it maintains consistency with the procedures used in the original publication on Study 4 (exp) [(Atlas et al., 2010)](https://paperpile.com/c/mO5B1U/4jqzz), and (b) it provides an opportunity to examine predictive performance using a flexible basis set. For both studies, the pain period basis set consisted of three curves shifted in time and was customized for thermal pain responses based on previous studies [(Atlas et al., 2010; Lindquist et al., 2009)](https://paperpile.com/c/mO5B1U/4jqzz+HPQBP). To estimate cue-evoked responses for Study 4 (exp), the pain anticipation period was modeled using a boxcar epoch convolved with a canonical HRF. This epoch was truncated at 8 s to ensure that fitted anticipatory responses were not affected by noxious stimulus-evoked activity. As with the other studies, we included nuisance covariates and excluded trials with VIFs > 2.5. In Study 4 (exp) we also excluded trials that were global outliers (those that exceeded 3 SDs above the mean). We reconstructed the fitted basis functions from the flexible single-trial approach to compute the area under the curve (AUC) for each trial and in each voxel. We used these trial-by-trial AUC values as estimates of trial-level anticipatory or pain-period activity.

**Computing Neurologic Pain Signature (NPS) responses**

For each trial and each subject, we computed a single scalar value representing the NPS pattern expression in response to the thermal and pressure pain stimulus (using the contrast [Pain Stimulation minus Baseline] images). There are primarily three methods to calculate the NPS pattern response, given the NPS is represented as a vector ***x***, brain response to pain stimulus as a vector ***y*** and the voxel number in the brain mask as ***n***: (1) dot-product (NPSdot = inxiyi), which combine magnitude and spatial similarity information; (2) correlation (NPScorr = in(xi-x)(yi-y)in(xi-x)2in(yi-y)2) which excludes information related to whole-image mean and scale, and represents spatial pattern similarity; (3) cosine similarity (NPScos = inxiyiinxi2inyi2), which is similar with correlation, but without mean-centering. The effect size and reliability of three versions of NPS responses were not significantly different from each other (see **Table S5**). Thus we reported the results of NPSdot in the main text.

To test whether NPS's performance exceeds individual brain regions within NPS, we also computed the pattern expression, i.e., dot-product, for each brain area within NPS. The local brain areas were extracted based on the NPS map thresholded at q < 0.05 FDR, k > 10. We compared the effect size and the reliability of individual brain regions with the whole NPS pattern. In most of the regions in the NPS, pain is associated with increased overall activity. Such regions include the major targets of ascending nociceptive afferents, including the thalamus, secondary somatosensory regions (S2), posterior, mid and anterior insula and adjacent opercula, dorsal anterior cingulate cortex (dACC). In a subset of other medial regions, including the perigenual ACC (pgACC) and the posterior cingulate (PCC)/precuneus/paracentral lobule, pain was associated with deactivation in the original NPS pattern. These regions are not strongly linked to nociception and are not direct targets of nociceptive afferents; rather, they have been associated with a variety of affective, autonomic, social, self-referential, and decision-making functions [(Roy et al., 2014)](https://paperpile.com/c/mO5B1U/4nE3q).

**Effect size analysis**

We analyzed four types of effect size of the NPS in the single-trial dataset: (1) Mean response: the mean NPS response across all trials irrespective of the temperature and experiment manipulations. One-sample t-test was conducted for all participants in each study; (2) within-person correlation with temperature: the standardized regression coefficient with temperature as the independent variable and NPS response as the dependent variable was estimated. A one-sample t-test was conducted for the regression coefficients of all participants for each study; (3) Within-person correlation with pain reports: the standardized regression coefficient with pain reports as the independent variable and NPS response as the dependent variable was estimated. A one-sample t-test was conducted for the regression coefficients of all participants for each study. (4) Between-person correlation with pain reports. The mean NPS response and mean pain reports of each participant were calculated by the average of each participant's trials. The correlation between the NPS response and pain reports was calculated across all participants for each study. The effect size was determined by Cohen's d values, which are commonly characterized as follows: 0.01 indicates very small; 0.20 indicates small; 0.50 indicates medium; 0.80 indicates large; 1.20 indicates very large, and 2.0 indicates a huge effect size [(Cohen, 2013; Sawilowsky, 2009)](https://paperpile.com/c/mO5B1U/IVZhP+wRjLX). In between-person correlations, the transformation between r and cohen's d is d = 2r1-r2.

**Test-retest reliability analysis**

Test-retest reliability of NPS and pain reports was determined by intra-class correlation coefficient (ICC; [Shrout and Fleiss 1979; McGraw and Wong 1996; Koo and Li 2016)](https://paperpile.com/c/mO5B1U/kyhz6+Q5sTB+Lm0S). We characterized two types of test-retest reliability, i.e., short-term and long-term test-retest reliability, based on the time interval between measures. In the single-trial dataset, which includes study 1 to 8, we calculated the short-term test-retest reliability since data were collected within one session. To do so, we constructed a two-way mixed effects model with time (1st vs. 2nd half trials) as a fixed effect, and subjects as a random effect. Since we were interested in the measure of the average of trials from the 1st and 2nd half of each study, the mixed-effect model is referred to as ICC(3,k) = (BMS - EMS) / BMS, where BMS represents between-person mean square, and EMS represents error mean square, i.e., within-person sum of squares minus between-time sum of squares. It is noteworthy that though the short-term test-retest reliability is mathematically identical to the internal consistency reliability, they are conceptually different. Internal consistency measures how consistent a set of items, e.g., voxels in NPS, measures a particular construct, e.g., pain [(Drost et al., 2011)](https://paperpile.com/c/mO5B1U/PbpGU). Whereas the short-term test-retest reliability characterizes the short-term temporal stability of a measurement, e.g., the NPS response measured within a session [(Drost et al., 2011)](https://paperpile.com/c/mO5B1U/PbpGU). High values of internal consistency are not alway desirable, and could point to redundancy of items [(Streiner, 2003)](https://paperpile.com/c/mO5B1U/g0Xl1), while high values of test-retest reliability are a desirable feature given the constructs being measured are stable.

For study 9 and 10, we calculated the long-term test-retest reliability since data were collected across multiple sessions with longer time intervals. We also constructed a two-way mixed effect model with time (multiple sessions) as a fixed effect, and subjects as a random effect. Since we are interested in the reliability of the measure in one session, instead of the average of all sessions, the mixed-effect model is referred to as ICC(3,1) = (BMS - EMS) / (BMS + (k - 1) \* EMS), where BMS represents between-person mean square, EMS represents error mean square, and k represents session number [(Shrout and Fleiss 1979; McGraw and Wong 1996; Koo and Li 2016)](https://paperpile.com/c/mO5B1U/kyhz6+Q5sTB+Lm0S). Measures with ICCs are commonly characterized as follows: less than .40 are thought to have poor reliability, between .40 and .60 fair reliability, .60 and .75 good reliability, and greater than .75 excellent reliability [(Cicchetti & Sparrow, 1981)](https://paperpile.com/c/mO5B1U/B5RiL). Besides the estimated ICC values, we also reported the 95% confidence interval of each ICC value [(McGraw and Wong 1996; Koo and Li 2016)](https://paperpile.com/c/mO5B1U/Q5sTB+Lm0S).

**Results**

**Four types of NPS effect size**

In the first single-trial dataset, mean responses of NPS were significantly larger than zero in each of 8 studies (t = 5.02 - 19.05, ps < 0.001, d = 1.22 - 2.62). The within-person regression coefficients between NPS and temperature were significantly larger than zero in each of 8 studies (mean beta = 0.05 - 0.42, t = 2.32 - 18.91, ps < 0.05, d = 0.53 - 2.74). The within-person regression coefficients between NPS and subjective pain reports were significantly larger than zero in each of 8 studies (mean beta = 0.14 - 0.35, t = 4.81 - 11.49, ps < 0.001, d = 0.94 - 2.13). Lastly, the between-person correlations between NPS and subjective pain rating (i.e., individual differences) were only significant in 1 out of 8 studies (pearson r = -0.13 - 0.74, p = 0.69e-3 - 0.70, d = -0.27 - 2.20; see **Figure 1** for four types of one-sample t statistics and the effect size; see **Table S1** for the statistical details of each study).

We did the same analyses for each local brain area of NPS and compared the effect sizes with NPS. Generally, positive brain regions had higher effect size than the negative brain regions, and the effect sizes of NPS were the highest in all four tests (see **Figure S1**). To confirm the difference of the effect sizes between NPS and local brain regions, we conducted paired t-tests treating study as the unit of the observation and controlled the false positive rate of multiple comparisons. NPS has significantly larger effect size than most local brain regions in the mean response, except for the rIns (mean±se = 1.92±0.16 vs. 1.72±0.19); has significantly larger effect size in the within-person correlation with the temperature, except for the rIns (1.50±0.27 vs. 1.21±0.26); has significantly larger effect size in the within-person correlation with the subjective pain reports, except for the dACC (1.45±0.16 vs. 1.19±0.12); but doesn’t have significantly different effect size in the between-person correlation with the subjective pain reports from most brain regions, except for the rIPL (0.49±0.26 vs. -0.27±0.17) (see **Table S2** for all statistic details).

**Test-retest reliability**

The short-term test-retest reliability of NPS calculated in the single-trial dataset was distributed from good to excellent among 9 studies (ICC = 0.73 - 0.91; mean±s.e. = 0.84±0.02; see **Table S3** for more details), which was significantly smaller than the reliability of subjective pain reports (ICC = 0.85 - 0.96; mean±s.e. = 0.92±0.01; paired-t test: t(7) = 4.11, p = 0.005). Reliability of NPS was the highest compared with local brain regions; reliability of rThal and pgACC were significantly smaller than reliability of NPS (q-fdr < 0.05; see **Figure 2(A)** and **Table S4** for statistical details).

*Most pain-predictive and reliable NPS regions*

Among the brain regions that are most represented in the NPS pattern, we identified several regions with the greatest promise for predicting stable individual differences in pain. Those regions ought to have the highest combination of pain prediction in both within- and between-person and reliability. This is because within-person correlation with pain reports is meaningful in terms of relationship between NPS and pain reports, which might be driven by factors separate from what drives interindividual differences. Between-person correlation with pain reports is of primary interest for stable individual differences, though the effect sizes are moderate in our results. Besides, reliability is an important precondition for predicting stable individual differences in pain. A combination of three cutoffs, i.e., d > 0.2 for the effect size of both within- and between-person correlation with pain, and ICC > 0.6 for short-term test-retest reliability, filtered out 6 local regions including bilateral insula, right dorsal posterior insula, dACC, right S2 and right Thalamus (see **Figure 2(B)** and **Table S2** and **Table S4**).

*Long time interval between sessions*

The long-term test-retest reliability was tested in study 9 (rtnf) and study 10 (olp4cbp). For study 9 (rtnf), both reliability of NPS and pain reports were excellent (ICC = 0.75, 95CI = [0.61, 0.85] and 0.86, 95CI = [0.78, 0.92]). Time interval between session 1 and session 2 was 5.15 ± 4.66 days, and the correlations of the NPS and pain reports between two sessions were high (r = .78 and .83, ps < 0.001). The time interval between session 2 and session 3 was 4.48 ± 2.59 days, and the correlations of the NPS and pain reports between two sessions were high (r = .78 and .88, ps < 0.001). The correlations of the NPS and pain reports between session 1 and session 3 were high as well (r = .71 and .87, ps < 0.001; see **Figure 2(C)**). For study 10 (olp4cbp), 40 chronic pain participants in the waiting list didn’t receive any therapy intervention between two sessions. In the waiting list group, the reliability of NPS was fair (ICC = 0.46, 95CI = [0.22, 0.65]; r = 0.47, p = 0.004; see Figure **2(D)**) and the reliability of pain reports was poor (ICC = 0.26, 95CI = [-0.15, 0.49]). Reliability of the 40 participants in the therapy group and 40 participants in the placebo group were poor (see **Table S3** for details).

*How does the number of trials influence reliability?*

We tested how the number of trials of the heat stimuli influence the test-retest reliability. The results in **Figure 3(A)** left panel showed that with more trials being averaged to calculate the NPS response, the higher ICC values in each of 8 studies. On average, 60 or more trials were required to achieve an excellent reliability of NPS. Given the same number of trials being averaged, ICC values in study 9 (rtnf; 30 trials) and study 10 (olp4cbp; 5 trials) with longer time intervals were comparable with ICC values of study 1 to 8. The trend was flatter for the test-retest reliability of subjective pain reports, which achieved ‘excellent’ level with even one trial.

*How does the effect size of stimuli influence reliability?*

The property of the stimulus itself might influence the reliability, such as the effect size it induced. For example, heat stimuli with higher temperatures might generally induce higher pain effects. The results in **Figure 3(B)** left panel showed that NPS responses induced by higher temperature had higher test-retest reliability. However, this was not the case for the subjective pain rating, which was very reliable across all temperatures. NPS responses might be more specific for high painful stimulus intensity, while subjective pain rating could represent a wider range of pain levels in a reliable way.

*How does the type of contrast influence reliability?*

There are two commonly used methods to calculate the brain response to an experimental condition, comparing a condition with the implicit baseline or to a control condition. The results in **Figure 3(C)** left panel showed that the reliability of NPS dropped when the response of NPS was calculated in contrast with a lower temperature, instead of the implicit baseline. The drop of the reliability was smaller in subjective pain reports. This finding indicates that using a contrast with a control condition with low reliability could reduce the reliability of the contrast measure.

**Discussion**

Current efforts towards the translation of brain biomarkers have renewed interest in brain measures' effect sizes and reliability. With a large effect size, a measure can be diagnostic of outcomes at the individual level [(Reddan, Lindquist & Wager, 2017; Poldrack, Baker & Durnez, 2017)](https://paperpile.com/c/mO5B1U/79XL3+K0Ve). Test-retest reliability is a prerequisite for stable prediction of individual differences [(Streiner, 2003; Drost et al., 2011; Nakagawa & Schielzeth, 2010; Bennett & Miller, 2010)](https://paperpile.com/c/mO5B1U/g0Xl1+PbpGU+kOBbu+UDwQ). We systematically evaluated the effect sizes and test-retest reliability of the NPS across ten studies and 441 participants. The NPS showed a very-large effect size in predicting within-person single-trial pain reports (d = 1.45) and a medium effect size in predicting individual differences in pain reports (d = 0.49). The NPS showed excellent short-term (within-day) test-retest reliability (ICC = 0.84). Reliability was comparable in two additional studies across 5-day (N = 27) and 1-month (N = 40) test-retest intervals.

The current study with large samples replicated the previous findings that the multivariate brain pattern had large effect sizes in predicting external variables at the within-person level ([Wager et al., 2013; Zunhammer et al., 2018; Chang et al., 2015; Krishnan et al., 2016; Geuter et al., 2020](https://paperpile.com/c/mO5B1U/EIYNO+i3j8M+zGR3c+xKdrz+4nfsj)). Interestingly, the effect sizes at the between-person level were much smaller. This inconsistency of the effect sizes could be led by the different sources of variance underlying the NPS responses and pain reports at different levels. At the within-person level, different temperatures across trials are among the primary sources leading to the NPS responses and pain reports' variances. The effect sizes of within-person correlations between the NPS and the stimuli temperature were distributed from medium to huge (Cohen's d = 0.53 - 2.67). The effect sizes of within-person correlations between the pain reports and the stimuli temperature were all distributed from very large to huge (Cohen's d = 1.58 - 12.41). Both the NPS and pain reports are responsive to noxious stimuli intensities.

However, at the between-person level, the NPS and pain reports' variances were driven by many factors that are irrelevant to the stimuli intensities (see **Figure 5(A)**). One person can report more pain than another because of differences in demographic variables, genetic factors, and psychosocial processes [(Woo and Wager 2016; Fillingim 2017)](https://paperpile.com/c/tXAZY0/tHM2+zTzM). For example, the individual difference of subjective pain reports might reflect communicative bias, such as "stoics" vs. "communicators." Meanwhile, the NPS responses might vary due to individual differences in task-related head movement [(Engelhardt et al., 2017)](https://paperpile.com/c/mO5B1U/BhXh5), respiration [(Chang & Glover, 2009; Power et al., 2019)](https://paperpile.com/c/mO5B1U/WRXzq+o4GsP), heart rate [(Chang, Cunningham & Glover, 2009)](https://paperpile.com/c/mO5B1U/JygmK), BOLD magnitude [(Levin et al., 2001)](https://paperpile.com/c/mO5B1U/SDkt0) and inter-individual variation in brain bases for pain reports [(Reddan & Wager, 2018)](https://paperpile.com/c/mO5B1U/Tcg89). The NPS responses and pain reports contain different sources of variances at the between-person level that decrease the correlation between two variables.

Both the NPS (ICC = 0.73 - 0.91) and pain reports (ICC = 0.85 - 0.96) showed excellent short-term (i.e., within one-day) test-retest reliability. Test-retest reliability of pain reports has been extensively examined in previous pain-related studies with ICCs ranging from 0.75 - 0.96 [(Letzen et al. 2014; Upadhyay et al. 2015; Letzen et al. 2016; Jackson et al. 2020)](https://paperpile.com/c/tXAZY0/OhJj+WNtH+VHyr+lmXh). Both our findings and previous studies showed the excellent test-retest reliability of pain reports in measuring individual differences in pain. Previous studies have also examined the test-retest reliability of the univariate brain measures to pain and showed widely varied ICCs in pain-related ROIs (0.32 - 0.88; [Letzen et al. 2014; Quiton et al. 2014; Upadhyay et al. 2015](https://paperpile.com/c/tXAZY0/OhJj+7Y5L+WNtH)), significantly activated clusters (0.33 - 0.74; [Jackson et al. 2020](https://paperpile.com/c/tXAZY0/lmXh)) and functional connectivities (-0.17 - 0.77; [Letzen et al. 2016](https://paperpile.com/c/tXAZY0/VHyr)). Compared with the previous univariate brain measures of pain, the current study examined a multivariate brain measure, the NPS, using a large dataset and showed consistently high performance of test-retest reliability.

To test whether the NPS measure is stable across longer time scales, we recruited another two studies with 5-days and one-month intervals between sessions. To test the long-term test-retest reliability could also help to clarify whether the NPS applied to evoked pain is a state or trait measure. State measures are expected to have high short-term reliability, but low long-term reliability; while trait measures should also have high long-term reliability [(](https://paperpile.com/c/mO5B1U/g0Xl1)[Spielberger, 1983](https://paperpile.com/c/mO5B1U/zqhJ6); [Streiner, 2003)](https://paperpile.com/c/mO5B1U/g0Xl1). We found that the NPS had high long-term reliability as well, which means that NPS is a more trait-like measure. But the main thing is that you need a lot of data per person (> 60 trials).

That means the two measures reflected stable individual differences across time. Compared with pain reports which were used in the translational scenarios often, the NPS was a reliable index in measuring pain and shared common variance in stimuli intensity, but also have distinct variances. These findings suggest a multimodal approach, using subjective reports and the NPS together, that might be useful to measure pain, which is similar to the multimodal diagnostic criteria used in many diseases, such as mood, eating, and sleeping disorders in depression [(Park et al., 2017)](https://paperpile.com/c/mO5B1U/mWw18).

The short-term (i.e., within one-day) test-retest reliability of NPS were distributed from good to excellent (ICC = 0.73 - 0.91) in eight studies.

Controlling for the number of trials averaged over in calculating the NPS response for an individual person, the longer time interval did not reduce the test-retest reliability of NPS in another two studies with 5-days and one-month intervals between sessions.

Time scale is interesting in two ways: (1) whether the NPS measure is stable across longer time scales; (2) whether the NPS applied to evoked pain is a state or trait measure. State is expected to have high short-term, but low long-term test-retest reliability, but trait should also have high longer term stability. We found that NPS has both high short-term and long-term reliability, which means that NPS is more trait-like. But the main thing is that you need a lot of data per person. Controlling for the number of trials averaged over in calculating the NPS response for an individual person, the longer time interval did not reduce the test-retest reliability of NPS in another two studies with 5-days and one-month intervals between sessions. A dynamic state measurement is sensitive to time interval when examining the test-retest reliability, as the state can vary across time [(Streiner, 2003)](https://paperpile.com/c/mO5B1U/g0Xl1). For example, state anxiety has short-term reliability around 0.9, but longer-term (on the order of 1 month) test-retest reliability of 0.1-0.4 [()](https://paperpile.com/c/mO5B1U/zqhJ6). At the same time, a measurement for a trait is less sensitive to time intervals. The high performance of NPS in both short-term and long-term test-retest reliability indicates that NPS reflects stable individual differences across time.

A multimodal approach, using subjective reports and the NPS together, might be useful to measure pain, which is similar to the multimodal diagnostic criteria used in many diseases, such as mood, eating, and sleeping disorders in depression [(Park et al., 2017)](https://paperpile.com/c/mO5B1U/mWw18).

Previous studies have shown that multivariate brain patterns developed by machine learning techniques have larger effect sizes than traditional voxelwise brain measures [(Wager et al., 2013; Zunhammer et al., 2018; Chang et al., 2015; Krishnan et al., 2016; Geuter et al., 2020)](https://paperpile.com/c/mO5B1U/EIYNO+i3j8M+zGR3c+xKdrz+4nfsj) and higher test-retest reliability [(Woo & Wager, 2016; Yoo et al., 2019; Kragel et al., 2020)](https://paperpile.com/c/mO5B1U/DkZcQ+O7M4q+R9WxY).

For both effect size and test-retest reliability, the complete NPS performance was better than constituent local brain regions. This finding is consistent with the argument that pain is encoded in distributed brain networks instead of a specific and isolated brain region [(Woo et al., 2017; Kragel et al., 2018; Reddan & Wager, 2018; Petre et al., 2020)](https://paperpile.com/c/mO5B1U/gxX3D+IrEW1+Tcg89+1MP3c). Interestingly, the performance of six local brain regions (i.e., bilateral insula, right dorsal posterior insula, dACC, right S2, and right thalamus) were consistently better than other brain regions in both effect size and test-retest reliability. The six high-performance brain regions largely overlapped with the classic 'pain-matrix' found in previous studies using the traditional univariate voxel-wise method [(Peyron, Laurent & García-Larrea, 2000; Wager et al., 2004; Lamm, Decety & Singer, 2011)](https://paperpile.com/c/mO5B1U/7Xe63+XMJnk+AKIG). This observation suggests a consensus between univariate and multivariate brain measures methods [(Haufe et al., 2014)](https://paperpile.com/c/mO5B1U/Q2pno). The increased effect sizes and reliability of the complete NPS suggests that other local brain regions in the NPS are meaningful in predicting individual differences in pain even though they might be less directly related to pain processes.

We tested several factors that might influence the test-retest reliability (see **Figure 4(B)**). Firstly, the measurement averaged with more trials is more reliable. Reliability could be attenuated by high error variance sourcing from trial-by-trial variance [(Chen et al., 2020)](https://paperpile.com/c/mO5B1U/DP9fF). By increasing the trial number, the reliability could be improved dramatically [(Dang, King & Inzlicht, 2020; Rouder & Haaf, 2019)](https://paperpile.com/c/mO5B1U/MeqM2+v53eR). In our estimation, more than 60 trials per condition are required to achieve excellent test-retest reliability, though this is rarely done in practice. Secondly, given the same number of trials in the experiment, the stimuli with a larger effect size, e.g., higher temperature, produce more reliable measurement. It might be because the NPS is more responsive to high pain stimuli than low pain stimuli. High-intensity stimuli are more salient for participants and increase the signal-to-noise ratio. Thirdly, the measure calculated in contrast with the baseline is more reliable than in contrast with a control condition, especially when the control condition's reliability is low, as was the case for lower temperature stimuli. Lastly, some operational changes across different sessions might reduce the test-retest reliability, such as habituation/sensation effect in pain perception [(Jepma, Jones & Wager, 2014)](https://paperpile.com/c/mO5B1U/PcEYe), interventions, and state instability across sessions. In study 10 (olp4cbp), the NPS reliability was lower in two groups of participants who received some chronic pain therapy than the participants on the waiting list. All these factors have a more extensive influence on the NPS than self-reported pain, further supporting the argument that NPS and pain reports contain different sources of variance.

The test-retest reliability characterized by ICC values relates to the ratio of the 'true' values' variability between subjects and the measurement error in observed measurements. If reliability is high, measurement errors are small compared to the true differences between subjects so that subjects can be relatively well distinguished based on the observed measurements. If measurement errors are large compared with the true differences between subjects, differences between subjects' measurements could be driven by error; thus, the reliability is low in this situation. Multivariate brain patterns with high test-retest reliability are suitable for measuring individual differences of the corresponding mental processes or clinical symptoms, such as prognostic biomarkers for future disease and predictive biomarkers for treatment response [(FDA-NIH Biomarker Working Group, 2016)](https://paperpile.com/c/mO5B1U/4Xnln). It is noteworthy that not all biomarkers are developed to measure individual differences. Some other biomarkers mainly focus on detecting variation within individuals, such as detecting states of pain, consciousness, disease that vary across time. It is more appropriate for those biomarkers to evaluate the within-person measurement error than the test-retest reliability [(Kragel et al., 2020)](https://paperpile.com/c/mO5B1U/R9WxY). Thus, it is necessary to examine different reliability measures according to the uses of the biomarkers.

Reliability is not a fixed property of a measurement or a measurement technology, such as fMRI; instead, it is a property of the scores on a measurement for a particular sample of participants [(Wilkinson et al., 1999; Streiner, 2003)](https://paperpile.com/c/mO5B1U/PN5zx+g0Xl1). Besides the measurement itself, reliability could be influenced by the sample attributes, such as the sample size and heterogeneity of samples—the more heterogeneous the sample, the larger variance between subjects, and the higher the measurement reliability [(Henson, Kogan & Vacha-Haase, 2001)](https://paperpile.com/c/mO5B1U/Av9bp). The data in the current study is quite diverse in the aspects of the noxious stimuli (such as temperature, duration, body location), experiment manipulation (such as expectation, self-regulation, neurofeedback), population (such different groups of healthy participants, and chronic pain patients), acquisition protocols (such as different scanners and parameters). We believe the effect size and test-retest reliability of NPS results in the current study are representative and generalizable. In the future, many other multivariate brain patterns need to be tested in a similar manner before they can be used as translational biomarkers. Our study provides a blueprint for future studies performing such reliability testing and suggests factors that could improve test-retest reliability in future measurements.

**References**

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, *145*, 137-165.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage*, *26*(3), 839-851.

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.

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.

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, *12*(1), 26-41.

Barnhart, H. X., Haber, M. J., & Lin, L. I. (2007). An overview on assessing agreement with continuous measurements. *Journal of biopharmaceutical statistics*, *17*(4), 529-569.

Bartoshuk, L. M., Duffy, V. B., Green, B. G., Hoffman, H. J., Ko, C. W., Lucchina, L. A., ... & Weiffenbach, J. M. (2004). Valid across-group comparisons with labeled scales: the gLMS versus magnitude matching. *Physiology & behavior*, *82*(1), 109-114.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage*, *37*(1), 90-101.

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging?. *Annals of the New York Academy of Sciences*, *1191*(1), 133-155.

Bennett, C. M., & Miller, M. B. (2013). fMRI reliability: influences of task and experimental design. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 690-702.

Chang, C., Cunningham, J. P., & Glover, G. H. (2009). Influence of heart rate on the BOLD signal: the cardiac response function. *Neuroimage*, *44*(3), 857-869.

Chang, C., & Glover, G. H. (2009). Relationship between respiration, end-tidal CO2, and BOLD signals in resting-state fMRI. *Neuroimage*, *47*(4), 1381-1393.

Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., & Wager, T. D. (2015). A sensitive and specific neural signature for picture-induced negative affect. *PLoS biology*, *13*(6), e1002180.

Chen, G., Padmala, S., Chen, Y., Taylor, P. A., Cox, R. W., & Pessoa, L. (2020). To pool or not to pool: Can we ignore cross-trial variability in FMRI?. *bioRxiv*.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American journal of mental deficiency*, [*86*(2), 127–137.](http://paperpile.com/b/mO5B1U/B5RiL)

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Cole, J. H., & Franke, K. (2017). Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in neurosciences*, *40*(12), 681-690.

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated?. *Trends in cognitive sciences*, *24*(4), 267-269.

DeYoung, C. G., Hirsh, J. B., Shane, M. S., Papademetris, X., Rajeevan, N., & Gray, J. R. (2010). Testing predictions from personality neuroscience: Brain structure and the big five. *Psychological science*, *21*(6), 820-828.

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, *463*(7281), 657-661.

Doyle, O. M., Mehta, M. A., & Brammer, M. J. (2015). The role of machine learning in neuroimaging for drug discovery and development. *Psychopharmacology*, *232*(21-22), 4179-4189.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and perspectives*, *38*(1), 105.

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., ... & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, 0956797620916786.

Engelhardt, L. E., Roe, M. A., Juranek, J., DeMaster, D., Harden, K. P., Tucker-Drob, E. M., & Church, J. A. (2017). Children’s head motion during fMRI tasks is heritable and stable over time. *Developmental cognitive neuroscience*, *25*, 58-68.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature methods*, *16*(1), 111-116.

FDA-NIH Biomarker Working Group. (2016). BEST (Biomarkers, EndpointS, and other Tools) resource.

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, (47), S102.

Geuter, S., Reynolds Losin, E. A., Roy, M., Atlas, L. Y., Schmidt, L., Krishnan, A., ... & Lindquist, M. A. (2020). Multiple Brain Networks Mediating Stimulus–Pain Relationships in Humans. *Cerebral Cortex*, *30*(7), 4204-4219.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, *5*, 13.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, *48*(1), 63-72.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, *87*, 96-110.

Haynes, J. D. (2015). A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron*, *87*(2), 257-270.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the teacher efficacy scale and related instruments. *Educational and psychological Measurement*, *61*(3), 404-420.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825-841.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical image analysis*, *5*(2), 143-156.

Jepma, M., Jones, M., & Wager, T. D. (2014). The dynamics of pain: evidence for simultaneous site-specific habituation and site-nonspecific sensitization in thermal pain. *The Journal of Pain*, *15*(7), 734-746.

Kragel, P. A., Koban, L., Barrett, L. F., & Wager, T. D. (2018). Representation, pattern information, and brain signatures: from neurons to neuroimaging. *Neuron*, *99*(2), 257-273.

Kragel, P., Han, X., Kraynak, T., Gianaros, P. J., & Wager, T. (2020). fMRI can be highly reliable, but it depends on what you measure.

Krishnan, A., Woo, C. W., Chang, L. J., Ruzic, L., Gu, X., Lopez-Sola, M., ... & Wager, T. D. (2016). Somatic and vicarious pain are represented by dissociable multivariate brain patterns. *Elife*, *5*, e15166.

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage*, *54*(3), 2492-2502.

Lanczos, C. (1964). Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, *1*(1), 76-85.

Levin, J. M., Frederick, B. D., Ross, M. H., Fox, J. F., von Rosenberg, H. L., Kaufman, M. J., ... & Renshaw, P. F. (2001). Influence of baseline hematocrit and hemodilution on BOLD fMRI activation. *Magnetic resonance imaging*, *19*(8), 1055-1062.

Lindquist, M. A., Loh, J. M., Atlas, L. Y., & Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *Neuroimage*, *45*(1), S187-S198.

López-Solà, M., Woo, C. W., Pujol, J., Deus, J., Harrison, B. J., Monfort, J., & Wager, T. D. (2017). Towards a neurophysiological signature for fibromyalgia. *Pain*, *158*(1), 34.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature neuroscience*, *14*(2), 154-162.

Ma, Y., Wang, C., Luo, S., Li, B., Wager, T. D., Zhang, W., ... & Han, S. (2016). Serotonin transporter polymorphism alters citalopram effects on human pain responses to physical pain. *Neuroimage*, *135*, 186-196.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, *1*(1), 30.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, *59*(3), 2636-2643.

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non‐Gaussian data: a practical guide for biologists. *Biological Reviews*, *85*(4), 935-956.

Noble, S., Scheinost, D., & Constable, R. T. (2019). A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *Neuroimage*, *203*, 116157.

Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1140-1152.

Pannunzi, M., Hindriks, R., Bettinardi, R. G., Wenger, E., Lisofsky, N., Martensson, J., ... & Deco, G. (2017). Resting-state fMRI correlations: from link-wise unreliability to whole brain stability. *Neuroimage*, *157*, 250-262.

Park, S. C., Kim, J. M., Jun, T. Y., Lee, M. S., Kim, J. B., Yim, H. W., & Park, Y. C. (2017). How many different symptom combinations fulfil the diagnostic criteria for major depressive disorder? Results from the CRESCEND study. *Nordic journal of psychiatry*, *71*(3), 217-222.

Petre, B., Kragel, P. A., Atlas, L. Y., Geuter, S., Jepma, M., Koban, L., ... & Wager, T. D. (2020). Evoked pain intensity representation is distributed across brain systems: A multistudy mega-analysis. *BioRxiv*.

Peyron, R., Laurent, B., & Garcia-Larrea, L. (2000). Functional imaging of brain responses to pain. A review and meta-analysis. *Neurophysiologie Clinique/Clinical Neurophysiology*, *30*(5), 263-288.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., ... & Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience*, *18*(2), 115.

Power, J. D., Lynch, C. J., Silver, B. M., Dubin, M. J., Martin, A., & Jones, R. M. (2019). Distinctions among real and apparent respiratory motions in human fMRI data. *NeuroImage*, *201*, 116041.

Reddan, M. C., Lindquist, M. A., & Wager, T. D. (2017). Effect size estimation in neuroimaging. *JAMA psychiatry*, *74*(3), 207-208.

Reddan, M. C., & Wager, T. D. (2018). Modeling pain using fMRI: from regions to biomarkers. *Neuroscience bulletin*, *34*(1), 208-215.

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: a robust approach. *Neuroimage*, *53*(4), 1181-1196.

Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage*, *23*(2), 752-763.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic bulletin & review*, *26*(2), 452-467.

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G. E., & Wager, T. D. (2014). Representation of aversive prediction errors in the human periaqueductal gray. *Nature neuroscience*, *17*(11), 1607-1612.

Saad, Z. S., Reynolds, R. C., Argall, B., Japee, S., & Cox, R. W. (2004, April). SUMA: an interface for surface-based intra-and inter-subject analysis with AFNI. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)* (pp. 1510-1513). IEEE.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 26.

Shmuel, A., Chaimow, D., Raddatz, G., Ugurbil, K., & Yacoub, E. (2010). Mechanisms underlying decoding at 7 T: ocular dominance columns, broad structures, and macroscopic blood vessels in V1 convey information on the stimulated eye. *Neuroimage*, *49*(3), 1957-1964.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Spielberger, C. D. (1983). State-trait anxiety inventory for adults.

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment*, *80*(1), 99-103.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging*, *29*(6), 1310-1320.

Wager, T. D., & Nichols, T. E. (2003). Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *Neuroimage*, *18*(2), 293-309.

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C. W., & Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*, *368*(15), 1388-1397.

Wager, T. D., Rilling, J. K., Smith, E. E., Sokolik, A., Casey, K. L., Davidson, R. J., ... & Cohen, J. D. (2004). Placebo-induced changes in FMRI in the anticipation and experience of pain. *Science*, *303*(5661), 1162-1167.

Wilkinson, L. (1999). Task Force on Statistical Inference. Statistical methods in psychology journals. *American Psychologist*, *54*(3), 594-604.

Woo, C. W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: brain models in translational neuroimaging. *Nature neuroscience*, *20*(3), 365.

Woo, C. W., & Wager, T. D. (2016). What reliability can and cannot tell us about pain report and pain neuroimaging. *Pain*, *157*(3), 511-513.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, *8*(8), 665-670.

Yoo, K., Rosenberg, M. D., Noble, S., Scheinost, D., Constable, R. T., & Chun, M. M. (2019). Multivariate approaches improve the reliability and validity of functional connectivity and prediction of individual behaviors. *Neuroimage*, *197*, 212-223.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, *20*(1), 45-57.

Zunhammer, M., Bingel, U., Wager, T. D., & Placebo Imaging Consortium. (2018). Placebo effects on the neurologic pain signature: a meta-analysis of individual participant functional magnetic resonance imaging data. *JAMA neurology*, *75*(11), 1321-1330.

Zuo, X. N., Di Martino, A., Kelly, C., Shehzad, Z. E., Gee, D. G., Klein, D. F., ... & Milham, M. P. (2010). The oscillating brain: complex and reliable. *Neuroimage*, *49*(2), 1432-1445.

Zuo, X. N., & Xing, X. X. (2014). Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective. *Neuroscience & Biobehavioral Reviews*, *45*, 100-118.

**Table 1.** Study demographics, experiment manipulations and prior publications

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Study** | **N** | **Gender** | **Ages, M (SD)** | **# of Sessions** | **Interval between Sessions (days)** | **Prior publications** |
| Study1  (bmrk3) | 33 healthy | 22 F | 27.9 (9.0) | 1 | N/A | [(Woo et al. 2015; Lindquist et al. 2017; Woo et al. 2017; Geuter et al. 2020)](https://paperpile.com/c/IsgZUW/fta3+2J0K+UgRC+j40x) |
| Study2  (bmrk4) | 28 healthy | 10 F | 25.2 (7.4) | 1 | N/A | [(Chang et al. 2015; Krishnan et al. 2016; Lindquist et al. 2017; Woo et al. 2017; Geuter et al. 2020)](https://paperpile.com/c/IsgZUW/uvdc+auiD+2J0K+UgRC+j40x) |
| Study3  (bmrk5) | 93 healthy | 49 F | 28.7 (5.7) | 1 | N/A | [(Losin et al. 2020; Geuter et al. 2020)](https://paperpile.com/c/IsgZUW/N20Q+j40x) |
| Study4  (exp) | 17 healthy | 9 F | 25.5 | 1 | N/A | [(Atlas et al. 2010; Lindquist et al. 2017; Woo et al. 2017; Geuter et al. 2020)](https://paperpile.com/c/IsgZUW/X2l3+2J0K+UgRC+j40x) |
| Study5  (ie) | 50 healthy | 27 F | 25.1 (6.9) | 1 | N/A | [(Roy et al. 2014; Lindquist et al. 2017; Woo et al. 2017; Geuter et al. 2020)](https://paperpile.com/c/IsgZUW/W3tz+2J0K+UgRC+j40x) |
| Study6  (ie2) | 19 healthy | 10 F | 25.5 (9.5) | 1 | N/A | [(Jepma et al. 2018; Geuter et al. 2020)](https://paperpile.com/c/IsgZUW/YWRv+j40x) |
| Study7  (ilcp) | 29 healthy | 16 F | 20.4 (3.3) | 1 | N/A | [(Lindquist et al. 2017; Woo et al. 2017)](https://paperpile.com/c/IsgZUW/2J0K+UgRC) |
| Study8  (scebl) | 26 healthy | 11 F | 28 (9.3) | 1 | N/A | [(Koban et al. 2019; Lindquist et al. 2017; Woo et al. 2017)](https://paperpile.com/c/IsgZUW/j4ob+2J0K+UgRC) |
| Study9  (rtnf) | 27 healthy | 16 F |  | 3 | Ses 1 to 2: 5.15 (4.66); Ses 2 to 3: 4.48 (2.59); | unpublished |
| Study10  (olp4cbp) | 120 chronic back pain | 61 F | 42.6 (15.6) | 2 | 25 - 40 | Ashar et al., submitted |

**Table 2**. Stimulation protocol

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Study** | **Stimulus**  **location** | **Stimulus Intensity (**ºC**)** | **Stimulus duration (seconds)** | **Trials per subject** | **Other experimental manipulations** |
| Study1 (bmrk3) | Arm | 44.3, 45.3, 46.3, 47.3, 48.3, 49.3 | 12.5 | 97 | Cognitive self-regulation intervention to increase or decrease pain |
| Study2 (bmrk4) | Arm, Foot | 46, 47, 48 | 11 | 81 | Combination of painful stimuli with heat-predictive visual cues for low, medium, and high pain |
| Study3 (bmrk5) | Arm | 47, 48, 49 | (8 and) 11 | 35 | none |
| Study4  (exp) | Arm | 41.1 - 47.1 | 10 | 64 | Combination of painful stimuli with heat-predictive auditory cues |
| Study5  (ie) | Arm | 46, 47, 48 | 11 | 48 | Combination of painful stimuli with heat-predictive visual cues and with a placebo manipulation |
| Study6  (ie2) | Leg | 48, 49 | 1.85 | 70 | Combination of painful stimuli with heat-predictive visual cues |
| Study7  (ilcp) | Arm | 44.7, 46.7 | 10 | 64 | Combination of painful stimuli with intervention for perceived control (making vs. observing cue choice) and expectancy (80% vs. 50% probabilities of low pain) |
| Study8  (scebl) | Leg | 48, 49, 50 | 1.85 | 96 | Combination of painful stimuli with heat-predictive visual cues and unreinforced social information |
| Study9  (rtnf) | Leg | 46, 47, 48 | 12 | 30 | Combine painful stimuli with neural feedback on suppressing NPS activity |
| Study10 (olp4cbp) | thumbnail | 4, 7 kg/cm2\* | 6 | 5 | Data collected in the context of a randomized controlled trial, including a psychotherapy treatment, placebo treatment, and treatment-as-usual control group |

\*Pressure pain in Study10 (olp4cbp)

Figure Captions

**Figure 1.** NPS pattern and effect tests. **(A)** **The multivariate brain pattern of NPS**. Ins denotes Insula, V1 primary visual area, S2 secondary somatosensory cortex, ACC anterior cingulate cortex, Thal thalamus, STS superior temporal sulcus, PCC posterior cingulate cortex, LOC lateral occipital complex, and IPL inferior parietal lobule. Direction is indicated with preceding lowercase letters as follows: r denotes right, l left, d dorsal, p posterior, pg perigenual. Surface data were visualized in SUMA [(Saad et al. 2004)](https://paperpile.com/c/mO5B1U/60vHR). **(B)** **Four types of NPS effect size and unfolded tests**. **Central panel**: Each big dot represents a type of averaged effect size of study 1 to 8; the vertical bar represents the standard error; each small dot represents the effect size of one study.**Top-left panel**: mean response of NPS. Each big dot represents the mean response of NPS in each study; the vertical bar represents the standard error; each small dot represents mean NPS response of one participant; and the violin plot represents the distribution of all participants in each study. To make the NPS response values comparable across different studies, the NPS response was rescaled by mean absolute deviation within each study. **Top-right panel**: within-person correlation between the NPS response and the temperature. Each big dot represents the mean beta value of the regression with the temperature as the independent variable and the NPS response as the dependent variable; the vertical bar represents the standard error; each small dot represents the beta value of one participant; and the violin plot represents the distribution of all participants in each study. **Bottom-left panel**: within-person correlation between the NPS response and the subjective pain reports. Each big dot represents the mean beta value of the regression with the subjective pain ratings as the independent variable and the NPS response as the dependent variable; the vertical bar represents the standard error; each small dot represents the beta value of one participant; and the violin plot represents the distribution of all participants in each study. **Bottom-right panel**: between-person correlation between the NPS response and participants’ mean subjective pain reports. Each dot represents one participant; the line represents the linear relationship between the mean of the subjective pain reports and the mean of the NPS response of each participant; and the shadow represents the standard error. The NPS response and pain ratings were rescaled by mean absolute deviation within each study. \*\*\* p < 0.001; \*\* p<0.005; \* p<0.05.

**Figure 2**. Test-retest reliability and high-ranked regions. **(A)** Short-term test-retest reliability of subjective pain reports, NPS and local regions. Each big dot represents the mean reliability of the study 1 to 8; the vertical bar represents the standard error; each small dot represents the reliability of one study. The downward-pointing arrows indicate ICC < 0. **(B)** Regions with greatest promise for predicting stable individual differences in pain: highest combination of pain prediction within and between person and short-term reliability. Cutoffs: reliability > 0.75, pain prediction > 0.2 for within and between person. **(C)** Illustration of long-term test-retest reliability with 1 week interval (ICC = 0.75). Correlations of NPS response among three sessions in study 9 (rtnf). **(D)** Illustration of long-term test-retest reliability with 1 month interval (ICC = 0.46). Correlations of NPS response among two sessions in study 10 (olp4cbp / waitlist group). Ins denotes Insula, V1 primary visual area, S2 secondary somatosensory cortex, ACC anterior cingulate cortex, Thal thalamus, STS superior temporal sulcus, PCC posterior cingulate cortex, LOC lateral occipital complex, and IPL inferior parietal lobule. Direction is indicated with preceding lowercase letters as follows: r denotes right, l left, d dorsal, p posterior, pg perigenual.

**Figure 3.** Factors that influence the reliability of the NPS response (left column) and subjective pain reports (right column). The small numbers from 1 to 8 correspond to the study 1 to 8. Study 9 is marked with the study name, i.e., ‘rtnf’; and study 10 is marked with the study name, i.e., ‘olp4cbp’. **(A)** Influence of the trial number and time interval between sessions. The short-term reliability of study 1 - 8 was calculated using ICC(3, k). The long-term reliability of study 9 and 10 was calculated using ICC(3,1). The ICC values were calculated based on different trial numbers. Each line with color shows the nonlinear relationship between the trial number and the ICC values of the corresponding study (fitted using the *loess* function in R). The ICC values estimated with less than 10 participants were excluded due to poor estimation. The black line showed the average of study 1 to 8, which was weighted by the square root of the number of participants in each study. The grey shadow presents the standard error, which was also weighted by the square root of the number of participants. On average, to achieve an excellent reliability, at least 60 trials were required to calculate the NPS response. Given the same number of trials, the long-term reliability of NPS in study 9 and 10 were comparable with the short-term reliability of NPS in study 1 to 8. Reliability of pain reports were excellent in general, but was poor in olp4cbp. **(B)** Influence of the temperature of the heat stimuli. The ICC values estimated with less than 13 participants with more than 4 trials in each temperature were excluded due to poor estimation. Under this criterion, the study 4 and 7 were with no ICC value presented in the plot. NPS responses are more reliable in higher temperature stimuli. Whereas pain reports are reliable across all temperature stimuli. **(C)** Influence of the single condition vs. contrast. The larger dots represent the ICC values calculated by a single temperature condition, and the smaller dots represent the ICC values calculated by the contrast of each temperature minus the lowest temperature in each study. The length of the dashed line represents the difference between the ICC values calculated by single temperature condition and by the contrast of two temperatures. The downward-pointing arrow indicates ICC < 0. The contrast is less reliable than the NPS response or pain report at a single temperature in virtually every case.

**Figure 4.** Summary and rationale for multi-modal measurement approach. **(A)** Different sources of error for different measured variables suggest a multi-modal approach to pain measurement is useful. Rectangles represent the observed variables, i.e., pain reports and NPS. Ellipses represent the latent variables that we aim to measure, i.e., the core nociceptive feeling. The circle represents sources of measurement error that add noise and bias to each observed measure. Both pain reports and NPS activity measure the core nociceptive circuits that generate pain experience. However, different sources of measurement error reduce the correlation between pain reports and NPS response. Using subjective and objective measures together could be a complementary measure for the core variable of interest. **(B)** Summary of factors that influence reliability. Rectangles represent the observed variables, such as NPS response, across different sessions. Ellipses represent the latent variables that we are interested to model. Results indicate that stimuli with larger effect size have higher test-retest reliability, indicated by the red plus ‘+’; some active change across sessions could decrease the test-retest reliability, indicated by the blue minus ‘-’. The circle represents the measurement error that could decrease the test-retest reliability, indicated by the blue minus ‘-’. Different width of the lines indicates active change might influence the measurement in different sessions to a different degree.

Figure 1.

Diagram, schematic

Description automatically generated

Figure 2.

Diagram

Description automatically generated

Figure 3.

Diagram

Description automatically generated

Figure 4.

Diagram

Description automatically generated