NeuroImage 247 (2022) 118844

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Effect sizes and test-retest reliability of the fMRI-based neurologic pain signature

Xiaochun Han[a,b], Yoni K. Ashar[c], Philip Kragel[d], Bogdan Petre[b], Victoria Schelkun[e], Lauren Y. Atlas[f,g,h], Luke J. Chang[b], Marieke Jepma[i], Leonie Koban[j], Elizabeth A. Reynolds Losin[k], Mathieu Roy[l], Choong-Wan Woo[m], Tor D. Wager[b,*]

[a] Faculty of Psychology, Beijing Normal University, Beijing, China
[b] Dartmouth College, Hanover, NH, United States
[c] Weill Cornell Medical College, New York, NY, United States
[d] Emory University, Atlanta, GA, United States
[e] Columbia University, New York, NY, United States
[f] National Center for Complementary and Integrative Health, National Institutes of Health, Bethesda, MD, United States
[g] National Institute of Mental Health, National Institutes of Health, Bethesda, MD, United States
[h] National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD, United States
[i] University of Amsterdam, Amsterdam, Netherlands
[j] INSEAD Fontainebleau & ICM paris, Paris, France
[k] Department of Psychology, University of Miami, Miami FL, United States
[l] Department of Psychology, McGill University, Montreal, Quebec, Canada
[m] Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, Gyeonggi-do, South Korea

ARTICLE INFO

Keywords:
Evoked pain
Individual differences
Measurement properties
Multivariate brain signature
Trial number

ABSTRACT

Identifying biomarkers that predict mental states with large effect sizes and high test-retest reliability is a growing priority for fMRI research. We examined a well-established multivariate brain measure that tracks pain induced by nociceptive input, the Neurologic Pain Signature (NPS). In $N = 295$ participants across eight studies, NPS responses showed a very large effect size in predicting within-person single-trial pain reports ($d = 1.45$) and 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, with average 69.5 trials/person). Reliability scaled with the number of trials within-person, with ≥60 trials required for excellent test-retest reliability. Reliability was tested in two additional studies across 5-day ($N = 29$, ICC = 0.74, 30 trials/person) and 1-month ($N = 40$, ICC = 0.46, 5 trials/person) test-retest intervals. The combination of strong within-person correlations and only modest between-person correlations between the NPS and pain reports indicate that the two measures have different sources of between-person variance. The NPS is not a surrogate for individual differences in pain reports but can serve as a reliable measure of pain-related physiology and mechanistic target for interventions.

1. Introduction

Understanding individual differences in brain activity and their links with behavior is a primary focus of fMRI research. One approach is to develop brain biomarkers that measure brain processes related to external constructs (e.g., pain, mental illness, other healthy and performance outcomes). Biomarkers can inform diagnosis and treatment, help subtype patient groups, or predict future risk of illness (FDA-NIH Biomarker Working Group, 2016). But to deliver on this potential, biomarkers must possess good measurement properties. In this paper, we assess the measurement properties of an evoked pain biomarker, the Neurological Pain Signature (NPS; Wager et al., 2013). We focused on two measurement properties: effect sizes in predicting pain and test-retest reliability. Large effect sizes indicate robust and replicable effects, and if they are sufficiently large, they may have sufficient precision to diagnose outcomes at the individual-person level (Poldrack et al., 2017; Reddan et al., 2017). Test-retest reliability is a prerequisite for prediction of stable individual differences (Bennett and Miller, 2010; Drost et al., 2011; Nakagawa and Schielzeth, 2010; Streiner, 2003).

Historically, the measurement properties of fMRI signals have been too infrequently assessed given their importance. Effect sizes for predicting external variables can be calculated at both within-person and between-person (individual differences) levels when repeated mea-

* Corresponding author at: 3 Maynard St., Hanover, NH 03784.
E-mail address: Tor.D.Wager@Dartmouth.edu (T.D. Wager).

https://doi.org/10.1016/j.neuroimage.2021.118844.
Received 12 July 2021; Received in revised form 13 December 2021; Accepted 19 December 2021
Available online 20 December 2021.

sures are collected from each person. Predictions at these levels can be inconsistent because they depend on different sources of variance (Bakdash and Marusich, 2017; Kievit et al., 2013). For example, two variables can be positively correlated at the within-person level but have no relation at the between-person level. Assessing effect sizes at both levels guards against incorrect interpretations of the predictions and facilitates a deeper understanding of the brain measures. Test-retest reliability reflects temporal stability under repeated tests and is usually measured with an intraclass correlation coefficient (ICC; Shrout and Fleiss, 1979). Both effect sizes and test-retest reliability rely on a low random error in the measurement. Test-retest reliability also relies on high inter-individual variability, indicating differentiable measures scores across subjects (Barnhart et al., 2007). Thus, quantifying effect size and test-retest reliability of fMRI-based biomarkers has the potential to greatly improve their utility and move the field towards more rigorous methods.

As translational goals accelerate and sample sizes increase, measurement properties of fMRI studies are increasingly a focus of attention (Bennett and Miller, 2010; Button et al., 2013; Dubois and Adolphs, 2016; Elliott et al., 2019, 2020; Hedge et al., 2018; Herting et al., 2018; Kraemer, 2014; Nichols et al., 2017; Noble et al., 2019; O'Connor et al., 2017; Poldrack et al., 2017; Xu et al., 2016; Zuo and Xing, 2014; Zuo et al., 2019). 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 et al., 2017; Marek et al., 2020). The reliability of univariate brain measures in many studies with small samples varies substantially (Letzen et al., 2016; Manuck et al., 2007; Nord et al., 2017; Plichta et al., 2012). 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). Similarly, univariate-style approaches to resting-state fMRI studies have found low test-retest reliability, with ICCs < 0.3 at the individual edge-level connectivity (Noble et al., 2019; Pannunzi et al., 2017).

However, alternatives to traditional single-region univariate analyses offer substantial promise. An important trend in the fMRI studies is the development of *a priori* multivariate brain measures that can be used as biomarkers, also called 'neuromarkers' or 'signatures' (Abraham et al., 2017; Arbabshirani et al., 2017; Doyle et al., 2015; Gabrieli et al., 2015; Haynes, 2015; Kragel et al., 2018; Orrù et al., 2012; Woo et al., 2017). 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 or participants. Because they are pre-specified models applied to new samples without re-fitting, neuromarkers provide an opportunity to evaluate measurement properties across different samples and contexts systematically. Multivariate brain signatures can yield measures with much larger effect sizes (Cohen's $d > 2$; Chang et al., 2015; Geuter et al., 2020; Krishnan et al., 2016; Marek et al., 2020; Wager et al., 2013; Zunhammer et al., 2018). They also show enhanced test-retest reliability for both task-evoked (ICCs > 0.7; Kragel et al., 2021; Woo and Wager, 2016) and resting-state (ICCs > 0.6; Gordon et al., 2017; Gratton et al., 2020; Yoo et al., 2019; Zuo and Xing, 2014) fMRI measures in some studies. However, this has rarely been assessed across diverse samples and scanners, particularly with respect to a systematic evaluation of effect sizes for within-person and between-person prediction of external variables and test-retest reliability.

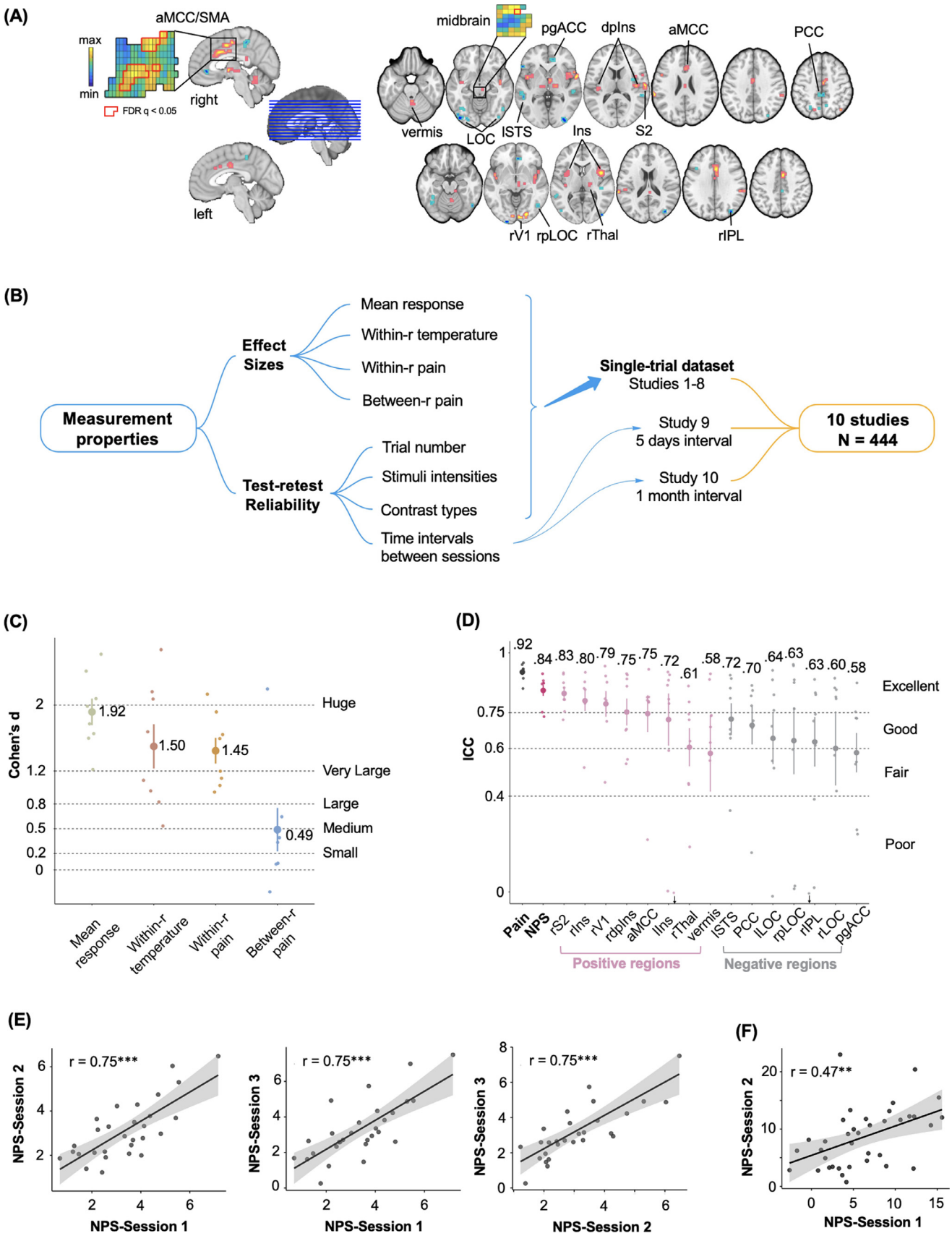
In the current study, we evaluated a well-established multivariate brain-based model related to pain , the Neurologic Pain Signature (NPS; Wager et al., 2013). The NPS was trained using machine-learning techniques to predict individual pain intensity ratings from brain activity during heat-induced pain in healthy participants. The training dataset included, for each participant, four trial-averaged pain-related activation maps and four trial-averaged pain reports for each of four stimulus intensities, including non-painful warmth, low pain, medium pain, and high pain calibrated for each participant to range from level 2 (barely painful) to level 8 (maximum pain participants would tolerate) on an 11-point visual analogue scale. The activation maps were calculated based on the contrast with the implicit baseline (i.e., the intercept in the general linear model estimation). By using activation maps from each pain intensity as a predictor set, and the averaged pain reports from each pain intensity as the outcome and using principal component regression to stabilize the parameter estimate maps, the algorithm provided interpretable brain maps composed of linear weights on voxels. The NPS consists of patterns within and across brain regions, including the thalamus, the posterior and middle insula, the secondary somatosensory cortex, the anterior cingulate cortex, the midbrain, and other regions (see Fig. 1(A)). The sensitivity and specificity of the NPS in response to painful stimuli were tested across different studies and samples. The NPS predicts subjective pain intensity in response to noxious thermal (Wager et al., 2013), mechanical (Krishnan et al., 2016), electrical (Krishnan et al., 2016; Ma et al., 2016), and visceral stimuli (Van Oudenhove et al., 2020). In addition, it does not respond to non-noxious warmth (Wager et al., 2013), threat cues (Krishnan et al., 2016; Ma et al., 2016; Wager et al., 2013), social rejection-related stimuli (Wager et al., 2013), vicarious pain (Krishnan et al., 2016), or aversive images (Chang et al., 2015), but may respond to some degree to certain somatomotor conditions (Harrison et al., 2021).

In addition, the NPS has been shown to have external validity in some clinical applications. (1) The NPS as a neuromarker linked to nociceptive pain can serve as an intermediate phenotype potentially relevant to 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). The NPS response to capsaicin-induced hyperalgesia has been found to be intact in healthy controls but absent in an individual with congenital insensitivity to pain caused by mutation in the $Na_v1.7$ gene (McDermott et al., 2019). (2) The NPS showed potential as a pharmacodynamic or response biomarker in some studies. For example, the magnitude of the NPS response to painful heat was substantially reduced when the analgesic agent remifentanil was administered (Wager et al., 2013). Following spinal manipulation, participants with neck pain showed decreased NPS activation (Weber et al., 2019). (3) The NPS response is largely unresponsive to placebo treatments, showing potential as a biomarker resistant to the placebo effects. A meta-analysis showed that placebo effects on the NPS were very small ($g = -0.07$). In comparison, placebo effects on the pain reports were moderate ($g = -0.66$) (Zunhammer et al., 2018).

NPS effect sizes have been mainly assessed on within-person correlations with pain (Lindquist et al., 2017) and reliability has only been assessed in a preliminary fashion (Kragel et al., 2021; Woo and Wager, 2016). The measurement properties of individual brain regions of the NPS have not been assessed systematically. Comparing the measurement properties of the whole NPS and individual brain regions could help clarify whether NPS's performance exceeds individual brain regions and reveal the different performances of different individual brain regions. Further, the properties that influence test-retest reliability of the NPS (e.g., amount of data collected per person) have 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.

2. Materials and methods

We tested four types of effect sizes and test-retest reliability for both the NPS and constituent local brain regions across ten studies (see Fig. 1(B); total $N = 444$). The effect sizes were tested in Studies


(A)
aMCC/SMA
max
min
FDR q < 0.05
right
left
midbrain
pgACC
dpIns
aMCC
PCC
vermis
LOC
lSTS
Ins
S2
rV1
rpLOC
rThal
rIPL
(B)
Measurement properties
Effect Sizes
Mean response
Within-r temperature
Within-r pain
Between-r pain
Test-retest Reliability
Trial number
Stimuli intensities
Contrast types
Time intervals between sessions
Single-trial dataset
Studies 1-8
Study 9
5 days interval
Study 10
1 month interval
10 studies
N = 444
(C)
Cohen's d
2
1.2
0.8
0.5
0.2
0
1.92
1.50
1.45
0.49
Huge
Very Large
Large
Medium
Small
Mean response
Within-r temperature
Within-r pain
Between-r pain
(D)
ICC
1
0.75
0.6
0.4
0
.92 .84 .83 .80 .79 .75 .75 .72 .61 .58 .72 .70 .64 .63 .63 .60 .58
Excellent
Good
Fair
Poor
Pain
NPS
rS2
rIns
rV1
rdpIns
aMCC
lIns
rThal
vermis
lSTS
PCC
lLOC
rpLOC
rIPL
rLOC
pgACC
Positive regions
Negative regions
(E)
NPS-Session 2
NPS-Session 1
r = 0.75***
NPS-Session 3
NPS-Session 1
r = 0.75***
NPS-Session 3
NPS-Session 2
r = 0.75***
(F)
NPS-Session 2
NPS-Session 1
r = 0.47**

1 to 8 (i.e., the single-trial dataset, see below for detailed description; $N = 295$), including mean response to painful stimuli, within-person correlation with temperature, within-person correlation with pain reports, and between-person correlation with pain reports. For test-retest reliability, we tested the short-term test-retest reliability of the NPS and local brain regions of interest in the single-trial dataset. With the single-trial dataset, we also examined several factors that might influence the performance of the test-retest reliability of the NPS, including the number of trials, stimuli intensities, and contrast types (Bennett and Miller, 2010, 2013). To further examine the impact of the time interval between sessions on the test-retest reliability, we included another two pain studies with five days (Study 9, $N = 29$) and one month (Study 10, $N = 120$) interval between sessions.

2.1. Data description

The single-trial dataset (i.e., Studies 1 to 8) included 15,940 single-trial images of fMRI activity from healthy subjects with multiple levels of noxious heat and pain ratings within one scan session (i.e., one day) across 295 participants. Participants received a series of painful stimuli and rated their individually experienced pain following each stimulus in all studies. Each study also included psychological manipulation (except for study 3), such as cue-induced expectation and placebo treatment. Study 9 collected behavioral and fMRI data from 29 healthy participants during heat pain tasks across three sessions with an average of five-day intervals between sessions. Study 10 collected behavioral and fMRI data from 120 participants with chronic back pain receiving pressure pain stimulation across two sessions with an average of one month between sessions. Descriptive data on age, sex, and other study sample features are given in Table 1. The number of trials, stimulation sites, stimulus intensities, and durations varied across studies but were comparable; these variables are summarized in Table 2. These studies were diverse in study-specific features and manipulations, which increased the generalizability of our findings.

Data from the Studies 1 to 8 have been used in previous publications (see Table 1). Other data have been published from Study 10, but the pressure-pain data reported here were not previously published. Data from study 9 were not previously published.All the analyses and findings reported here are novel, and the data used for developing the NPS (i.e., the data in the first study of Wager et al., 2013) were not included in the current study to avoid double-dipping (Kriegeskorte et al., 2009). 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).

2.2. 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, except for study 3 using the Pathway ATS model and study 8 with a 32 mm Peltier thermode endplate. Study 10 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). 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 s. Most studies applied thermal stimulation to the left volar forearm; study 2 also applied the left foot's dorsum; study 6 and study 8 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.

2.3. fMRI preprocessing

We maintained the preprocessing pipelines from the original published studies despite variations across studies as this will likely reflect the variations in preprocessing steps observed across studies in the literature. In studies 1 to 8, structural T1-weighted images were coregistered to each subject's mean functional image using the iterative mutual information-based algorithm implemented in SPM (Ashburner and Friston, 2005). They were then normalized to MNI space using SPM. Following SPM normalization, study 4 included an additional step of normalization to the group mean using a genetic algorithm-based normalization (Atlas et al., 2010, 2014; Wager and Nichols, 2003). 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

Fig. 1. NPS pattern and measurement properties. **(A) NPS pattern weights**. The map shows thresholded voxel weights at $q<0.05$ false discovery rate (FDR) for display only; all weights were used in the subsequent analyses. Two examples of aMCC/SMA and midbrain unthresholded patterns are presented in the insets; small squares indicate individual voxel weights. Ins denotes Insula, V1 primary visual area, S2 secondary somatosensory cortex, MCC midcingulate 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, m middle, d dorsal, p posterior, pg perigenual. **(B) A diagram summarizing analyses of measurement properties and the corresponding data used in different studies.** Measurement properties include effect sizes in predicting external variables (temperature and pain reports) and test-retest reliability of the NPS. We analyzed four types of effect sizes, including: (1) the mean response of the NPS to stimulation; (2) the within-person correlation between the NPS response and the temperature of the heat stimuli; (3) the within-person correlation between the NPS response and pain ratings; and (4) the between-person correlation between the NPS response and pain ratings. Besides assessing the test-retest reliability of the NPS, we also analyzed four factors that might influence the test-retest reliability, including the number of trials, stimuli intensity, contrast types (i.e., baseline condition), and the time interval between sessions. The influence of the time intervals between sessions was assessed with the data in Studies 9 and 10. All other analyses of the effect sizes and test-retest reliabilities were assessed with the data in Studies 1 to 8 (i.e., the single-trial dataset). **(C) Four types of NPS effect size**. Each big dot represents a type of averaged effect size of studies 1 to 8. The vertical bar represents the standard error. Each small dot represents the effect size of one study. See Figure S1 for the results for each study. See Figure S2 for the effect sizes of local regions of the NPS. **(D) Short-term test-retest reliability of subjective pain reports, NPS, and local regions.** Each big dot represents the mean reliability of studies 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 (presumably due to noise). See Figure S3 for the illustration of short-term test-retest reliability of the NPS and subjective pain reports. **(E) Illustration of longer-term test-retest reliability of NPS with a 5-day interval**. Correlations of the NPS responses between session 1, session 2 and session 3 in study 9 (ICC = 0.73). Each dot represents one participant. The line represents the linear relationship between the NPS response in sessions 1, 2 and 3, and the shadow represents the standard error. **(F) Illustration of longer-term test-retest reliability of NPS with a 1-month interval**. Correlation of the NPS responses between session 1 and session 2 in the treatment-as-usual control group of study 10 (ICC = 0.46). Note that this test involved fewer trials per participant, which also limits reliability. Each dot represents one participant. The line represents the linear relationship between the NPS response in sessions 1 and 2, and the shadow represents the standard error. $^{***}\ p < 0.001$; $^{**}\ p < 0.005$.

Table 1
Study demographics, experiment sessions and prior publications.

Study	N	Gender	Ages, M (SD)	# of Sessions	Interval between Sessions (days)	Prior publications
Study1	33 healthy	22 F	27.9 (9.0)	1	N/A	(Geuter et al., 2020; Lindquist et al., 2017; Woo et al., 2015; Woo et al., 2017)
Study2	28 healthy	10 F	25.2 (7.4)	1	N/A	(Chang et al., 2015; Geuter et al., 2020; Krishnan et al., 2016; Lindquist et al., 2017; Woo et al., 2017)
Study3	93 healthy	49 F	28.7 (5.7)	1	N/A	(Geuter et al., 2020; Losin et al., 2020)
Study4	17 healthy	9 F	25.5	1	N/A	(Atlas et al., 2010; Geuter et al., 2020; Lindquist et al., 2017; Woo et al., 2017)
Study5	50 healthy	27 F	25.1 (6.9)	1	N/A	(Geuter et al., 2020; Lindquist et al., 2017; Roy et al., 2014; Woo et al., 2017)
Study6	19 healthy	10 F	25.5 (9.5)	1	N/A	(Geuter et al., 2020; Jepma et al., 2018)
Study7	29 healthy	16 F	20.4 (3.3)	1	N/A	(Lindquist et al., 2017; Woo et al., 2017)
Study8	26 healthy	11 F	28 (9.3)	1	N/A	(Koban et al., 2019; Lindquist et al., 2017; Woo et al., 2017)
Study9	29 healthy	16 F	29.9 (9.8)	3	Ses 1 to 2: 4.93 (4.57); Ses 2 to 3: 4.79 (2.81);	unpublished
Study10	120 chronic back pain	61 F	42.6 (15.6)	2	25 - 40	(Ashar et al., 2021)

Table 2
Stimulation protocol.

Study	Stimulus location	Stimulus Intensity (˚C)	Stimulus duration (seconds)	Trials per subject	Other experimental manipulations
Study1	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	Arm, Foot	46, 47, 48	11	81	Combination of painful stimuli with heat-predictive visual cues for low, medium, and high pain
Study3	Arm	47, 48, 49	8 and 11	36	Heat stimuli were intermixed with physically and emotionally aversive sound stimuli
Study4	Arm	41.1 - 47.1	10	64	Combination of painful stimuli with heat-predictive auditory cues
Study5	Arm	46, 47, 48	11	48	Combination of painful stimuli with heat-predictive visual cues and with a placebo manipulation
Study6	Leg	48, 49	1.85	70	Combination of painful stimuli with heat-predictive visual cues
Study7	Arm	43.5 - 47.5	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	Leg	48, 49, 50	1.85	96	Combination of painful stimuli with heat-predictive visual cues and unreinforced social information
Study9	Leg	46, 47, 48	12	30	Combine painful stimuli with neural feedback on suppressing NPS activity
Study10	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

* Study 10 delivered pressure rather than thermal stimulation.

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 (overtime) were computed. Any values with a significant $\chi 2$ 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 $2 \times 2 \times 2$ mm^3 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).

The preprocessing of study 9 and 10 were conducted using *fMRIPrep* 1.2.4 (Esteban et al., 2019). The BOLD reference was co-registered to the T1w reference. 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. The BOLD time-series were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. The BOLD time-series were resampled to *MNI152NLin2009cAsym* standard space, generating a *preprocessed BOLD* run in *MNI152NLin2009cAsym* space. The preprocessed BOLD runs were smoothed with a 6 mm FWHM Gaussian kernel. We identified image-intensity outliers (i.e., 'spikes') using Mahalanobis distances (3 standard deviations) and dummy regressors were included as nuisance covariates in the first level. Twenty-four head motion covariates per run were entered into the first level model as well (displacement in six dimensions, displacement squared, derivatives of displacement, and derivatives squared).

2.4. General linear model (GLM) analyses

For studies 1 to 8, 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 et al., 2004). 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. Single-trial analysis for study 2 and 4 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 (Atlas et al., 2010), 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). To estimate cue-evoked responses for study 4, 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 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. For studies 9 and 10, we estimated a GLM for each participant, including the nuisance covariates generated in preprocessing and three regressors of interest: pain stimuli, pain ratings, and button presses, each convolved with the standard HRF.

2.5. Computing neurologic pain signature (NPS) responses

We computed a single scalar value for each trial and each subject, representing the NPS pattern expression in response to the thermal and pressure pain stimulus (using the contrast [Pain Stimulation minus Baseline] images). There are three methods to calculate the NPS pattern response, given the NPS is represented as a vector $\boldsymbol{x}$, brain response to pain stimulus as a vector $\boldsymbol{y}$, and the voxel number in the brain mask as $\boldsymbol{n}$: (1) dot-product ($\text{NPS} = \sum_i^n x_i y_i$), which combine whole-image magnitude and spatial similarity information; (2) cosine similarity ($\text{NPScos} = \frac{\sum_i^n x_i y_i}{\sqrt{\sum_i^n x_i^2}\sqrt{\sum_i^n y_i^2}}$), which excludes whole-image magnitude information, representing the dot-product of unit vectors; (3) correlation ($\text{NPScorr} = \frac{\sum_i^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i^n (x_i-\bar{x})^2}\sqrt{\sum_i^n (y_i-\bar{y})^2}}$), which excludes information related to whole-image mean and magnitude, equivalent to the cosine similarity between centered vectors. The effect size and reliability of these three NPS response metrics were not significantly different from each other (see Table S5). We reported the results of the dot-product of the NPS in the main text, considering it is the most often reported metric in the published papers examining the NPS and facilitating the comparison with other studies. The dot product metric (unlike cosine similarity and correlation) includes signal related to overall image intensity, which may be a feature of intense stimuli that activate diffuse neuromodulatory system (Lohani et al., 2017).

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 individual brain areas were defined based on the NPS map thresholded at q < 0.05 FDR, and k > 10 contiguous voxels. Firstly, we identified the peak voxel in each cluster surviving correction. Secondly, we applied a Gaussian smoothing kernel of 4-mm FWHM around each peak voxel to generate a mask that included all voxels defining the local pattern for each region. Then, we applied the mask to the original NPS pattern (i.e., unthresholded NPS map including all voxel weights). See López-Solà et al., 2017 for a more detailed description method of the local region definition. We compared the effect size and the reliability of individual brain regions with the whole NPS pattern using paired t-tests by treating the study as the unit of observation and corrected for multiple comparisons using $q < 0.05$ FDR. In most of the regions in the NPS, pain is associated with the increased overall activity, i.e., positive brain regions, including the right middle Insula (rmIns), the right dosal posterior Insula (rdpIns), the left middle Insula (lmIns), the right secondary somatosensory cortex (rS2), the anterior midcingulate cortex (aMCC), the right Thalamus (rThal), vermis and the right primary visual area (rV1). Such regions include the major targets of ascending nociceptive afferents. In a subset of other regions, pain is associated with the decreased overall activity, i.e., negative brain regions, including the perigenual ACC (pgACC), the posterior cingulate cortex (PCC), right inferior parietal lobule (rIPL), left lateral occipital complex (lLOC), right posterior lateral occipital complex (rpLOC), right lateral occipital complex (rLOC), and left superior temporal sulcus (lSTS). 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., 2012, 2014).

2.6. Effect size analysis

We analyzed four types of effect sizes of the NPS in the single-trial dataset. (1) *Mean response [Pain minus Baseline]*: the mean NPS response across all trials irrespective of the temperature and experiment manipulations. A one-sample *t*-test was conducted for all participants in each study. (2) *within-person correlation with temperature*: correlation between the temperature and NPS response. A one-sample *t*-test was conducted for the correlation coefficients of all participants for each study. (3) *Within-person correlation with pain reports*: correlation between pain reports and the NPS response. A one-sample *t*-test was conducted for the correlation 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: 0.20 indicates small; 0.50 indicates medium; 0.80 indicates large, and 1.20 indicates very large effect size (Cohen, 2013; Sawilowsky, 2009). In between-person correlations, the transformation between r and cohen's d is $d = \frac{2r}{\sqrt{1-r^2}}$.

2.7. Test-retest reliability analysis

Test-retest reliability of the mean NPS response [Pain minus Baseline] was determined by the intra-class correlation coefficient (ICC; Koo and Li, 2016; McGraw and Wong, 1996; Shrout and Fleiss, 1979). To compare with the NPS, we also tested the reliability of the mean pain reports using ICC. As an index to characterize the temporal stability of individual differences (i.e., between-participant reliability), a large ICC requires both high inter-individual variability and low intra-individual variability. High inter-individual variability implies highly differentiable measures across subjects, and low intra-individual variability indicates high stability across different time points (Barnhart et al., 2007). ICC is calculated by mean squares obtained through analysis of variance among a given set of measures. We characterized two types of test-retest reliability, i.e., short-term and longer-term test-retest reliability, based on the time interval between measures. In the single-trial dataset, which includes studies 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 of the trials) as a fixed effect and subjects as a random effect. Since we were interested in the reliability of the averaged measures of the 1st and 2nd half of the trials (i.e., the average of two halves, $k = 2$), the mixed-effect model is referred to as ICC(3,k) = (BMS - EMS) / BMS. BMS represents the mean square for between-person measures, and EMS represents the mean square for error. The ICC values in the current study were calculated using the ICC function in the 'psych' library in R.

For studies 9 and 10, we assessed the longer-term test-retest reliability since data were collected across sessions with longer time intervals. We also constructed a two-way mixed-effects model with time (multiple sessions) as a fixed effect and subjects as a random effect. Instead of calculating ICC(3,k), we calculated ICC(3,1) = (BMS - EMS) / (BMS + (k - 1) * EMS) for longer-term test-retest reliability since we were interested in the measure of one session, not the average of all sessions. BMS represents the mean square for between-person measures, EMS represents the mean square for error, and k represents the number of scanning sessions (Koo and Li, 2016; McGraw and Wong, 1996; Shrout and Fleiss, 1979). Measures with ICCs are commonly characterized as follows: less than 0.40 are thought to have poor reliability, between 0.40 and 0.60 fair reliability, 0.60 and 0.75 good reliability, and greater than 0.75 excellent reliability (Cicchetti and Sparrow, 1981). We also reported the 95% confidence interval of ICC values (Koo and Li, 2016; McGraw and Wong, 1996).

3. Results

3.1. NPS effect sizes

We tested four types of effect sizes of the NPS in the single-trial dataset. (1) *Mean response [Pain minus Baseline]*: mean responses of the NPS were significantly larger than zero in each of the 8 studies ($t = 5.02$ - 19.22, ps < 0.001; mean $d = 1.92$, ranging from 1.22 to 2.62). (2) *Within-person correlation with temperature*: the within-person correlations between the NPS and temperature were significantly larger than zero in each of the 8 studies as well (mean $r = 0.05$ - 0.42, $t = 2.32$ - 18.91, ps < 0.05; mean $d = 1.50$, ranging from 0.53 to 2.67). (3) *Within-person correlation with pain reports*: the within-person correlations between the NPS and subjective pain reports were significantly larger than zero in each of the 8 studies (mean $r = 0.14$ - 0.35, $t = 4.81$ - 11.49, ps < 0.001; mean $d = 1.45$, ranging from 0.94 - 2.13). (4) *Between-person correlation with pain reports*: the between-person correlations between the mean NPS and mean subjective pain rating (i.e., individual differences) were only significant in 1 out of 8 studies ($r = -0.13$ - 0.74, $p = 0.0007$ - 0.70; mean $d = 0.49$, ranging from −0.27 to 2.20; see Fig. 1 and Figure S1 for four types of tests and effect sizes; see Table S1 for the statistical details of each study).

In study 4, the only individual study that showed a significant between-person correlation, the stimuli were tailored to the individuals to elicit matched subjective pain. Except for studies 4 and 7, other studies applied the same temperatures to all participants, and the individual differences in subjective pain were not stimulus-driven. The participant sample size of study 4 was the smallest ($N = 17$) among studies 1 to 8. Recent studies have shown inflated between-subject effect size statistics in small sample size studies (Marek et al., 2020). In these studies, the between-person effect sizes were not significantly correlated with the number of participants per study ($r = -0.41$; $p = 0.32$). Larger between-person effect sizes might also be associated with the number of trials per participant. More data per participant could reduce within-person variance around each person's true value, reducing error in the between-person correlation. However, here, between-person effect sizes were not significantly correlated with the number of trials per participant ($r = 0.25$; $p = 0.56$).

To test whether NPS's performance exceeds individual brain regions within the NPS, we did the same analyses for each local brain area of the NPS and compared the effect sizes with the NPS. Generally, positive brain regions had higher effect sizes than negative brain regions and the effect sizes of the full NPS were the highest in all four tests (see Figure S2). To confirm the difference in the effect sizes between NPS and local brain regions, we conducted paired t-tests treating the study as the unit of the observation and corrected the multiple comparisons using $q < 0.05$ FDR. The NPS has (1) significantly larger effect size than most local brain regions in the mean response, except for the rmIns (NPS vs. the local region mean±se = 1.92±0.16 vs. 1.72±0.19); (2) significantly larger effect size in the within-person correlation with the temperature, except for the rmIns (1.50±0.27 vs. 1.21±0.26); (3) significantly larger effect size in the within-person correlation with the subjective pain reports, except for the aMCC (1.45±0.16 vs. 1.19±0.12); (4) does not significantly differ in 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).

3.2. Test-retest reliability

The short-term test-retest reliability of the NPS calculated in the single-trial dataset was distributed from good to excellent among the 8 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 the NPS was numerically higher than any local brain regions and was significantly higher than rThal and pgACC ($q < 0.05$ FDR; see Fig. 1(D) and Table S4 for statistical details).

3.3. NPS local regions with good measurement properties

While brain-wide measures like the NPS have favorable measurement properties compared with local signals, there is still much value in examining local regions. For example, in Wager et al., 2013, the local pattern within dpIns could distinguish between somatic (pain) and non-somatic (rejection) stimuli. The dpIns is also the first cortical target of ascending nociceptive pathways through VMpo thalamus (Craig, 2006), and contains more body site-specific information (Krishnan et al., 2016), and can trigger pain when stimulated (Mazzola et al., 2012). In subsequent work, the local NPS pattern within dpIns was the only region to distinguish pain from breathlessness and somatomotor stimulation (Harrison et al., 2021). Conversely, the local NPS pattern within aMCC better distinguished aversive (pain or rejection) from non-aversive stimuli (non-painful warmth). We identified several local regions of the NPS with relatively good measurement properties. A combination of three cutoffs, i.e., $d > 0.2$ for both within-person and between-person correlation with pain, and ICC > 0.6 for short-term test-retest reliability, identified six reliable local-region patterns: lmIns, rmIns, rdpIns, aMCC, rS2, and rThal (see Table S2 and Table S4). These local regions have relatively better measurement properties than other local regions, such as rV1, vermis, and NPS regions with predominantly negative weights (e.g., pgACC).

3.4. Longer-term test-retest reliability

The longer-term test-retest reliability was tested in studies 9 and 10. For study 9, both reliability of the NPS and pain reports were excellent (ICC = 0.74, 95CI = [0.61, 0.84] and 0.87, 95CI = [0.80, 0.92]; see Fig. 1(E)). The time interval between session 1 and session 2 was 4.93 ± 4.57 days, and the time interval between session 2 and session 3 was 4.79 ± 2.81 days. Study 10 was a clinical trial randomizing chronic back patients to a psychological treatment, a placebo treatment, or a control group ($n = 40$ per group), with approximately 1 month between the two assessment sessions. In the control group, the reliability of the NPS was fair (ICC = 0.46, 95CI = [0.22, 0.65]; see Fig. 1(F)) and the reliability of pain reports was poor (ICC = 0.26, 95CI = [−0.15, 0.49]). The reliabilities of the NPS and pain reports in the psychotherapy group

and the placebo group were poor (see Table S3 for details). Reliability in study 10 was likely limited by the low number of trials (5 trials per person) used in this study (see next section).

3.5. How does the number of trials influence reliability?

We tested how the number of trials of the heat stimuli influences the test-retest reliability. The results in Fig. 2(A) left panel showed that the more trials averaged to calculate the NPS response, the higher the ICC values in each of the 8 studies. On average, 60 or more trials per condition were required to achieve excellent reliability of the NPS. Given the same number of trials being averaged, ICC values in study 9 (30 trials) and study 10 (5 trials) with longer time intervals were comparable with ICC values of studies 1 to 8. The trend was flatter for the test-retest reliability of subjective pain reports, which achieved an excellent level with even one trial.

3.6. 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 Fig. 2(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.

3.7. 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 Fig. 2(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 (ICC mean±*s*.e. = 0.25±0.17 vs. 0.81±0.03, which was calculated by averaging the reliability of all temperatures in one study first and calculating the mean and standard error of the reliability across all studies. Same below.). The drop of the reliability was smaller in subjective pain reports (ICC mean±*s*.e. = 0.80±0.03 vs. 0.93±0.01). This finding indicates that using a contrast with a control condition with low reliability could reduce the reliability of the contrast measure.

4. Discussion

Identifying biomarkers with good measurement properties is a growing priority for fMRI research. In the current paper, we systematically evaluated the effect sizes and the test-retest reliability of the NPS across ten studies and 444 participants. The NPS showed a very large effect size in predicting within-person single-trial pain reports (mean $d = 1.45$, ranging from 0.94 to 2.13). The effect size in predicting individual differences in pain reports is medium and heterogeneous across studies (mean $d = 0.49$, ranging from −0.27 to 2.20, equivalent to $r = 0.20$). The NPS showed excellent short-term (within-day) test-retest reliability (mean ICC = 0.84). Reliability was comparable in a study with a longer time interval across 5-day ($N = 29$, ICC = 0.74). It was lower in a study with 1-month test-retest intervals ($N = 40$, ICC = 0.46), though this may have been driven by the low number of trials (5 trials per person) rather than the longer time interval between sessions.

The current findings with a large sample of participants indicate that the NPS measures neurophysiological processes related to evoked pain with large effect sizes at the within-person level and high test-retest reliability. However, as a measure of individual differences in pain sensitivity, the NPS is only modestly related to the pain reports. This inconsistency of the effect sizes at within-person and between-person levels

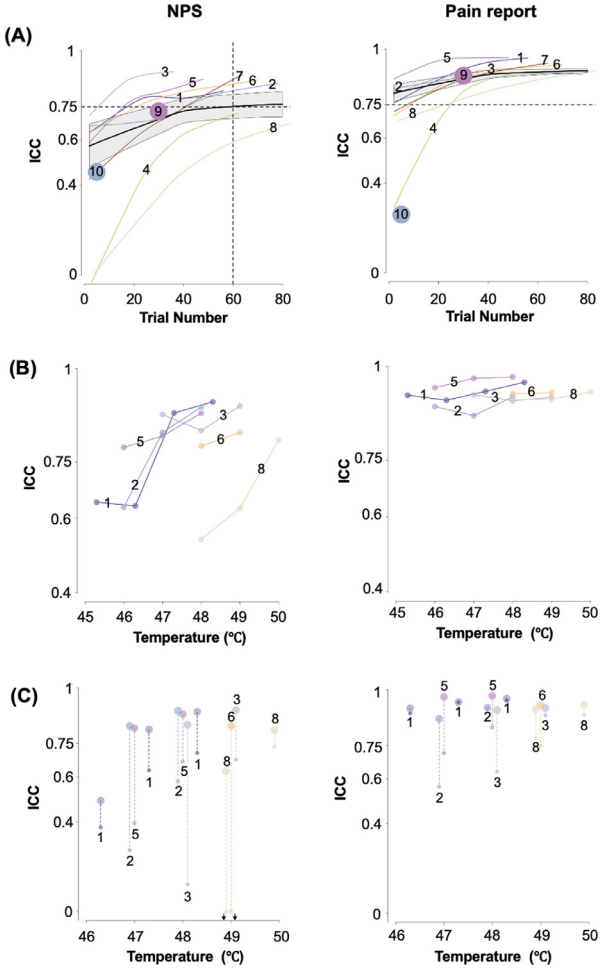


Fig. 2. Factors that influence the reliability of the NPS response (left column) and subjective pain reports (right column). The small numbers from 1 to 10 correspond to studies 1 to 10. **(A) The Influence of the trial number and time interval between sessions.** The ICC values were calculated based 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 studies 1 to 8, which was weighted by the square root of the number of participants in each study. The gray shadow presents the standard error, which was also weighted by the square root of the number of participants in each study. On average, to achieve excellent reliability, at least 60 trials were required to calculate the NPS response. Reliability was comparable in studies 9 and 10 with a longer time interval across 5-day and 1-month given the same number of trials (trial number = 30 and 5). The reliability of pain reports were excellent in general but were poor in study 10. **(B) The influence of the temperature of the heat stimuli.** Only participants with more than 4 trials in each temperature were included in the ICC calculation. The ICC values estimated with less than 13 participants were excluded due to poor estimation. Under these criteria, 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) The influence of the types of contrast.** The larger dots represent the ICC values of the measurements calculated by comparing a temperature condition with the baseline, and the smaller dots represent the ICC values of the measurements calculated by comparing a temperature condition with the lowest temperature condition in each study. The length of the dashed line represents the difference between the ICC values of measurements calculated with different types of contrast. The downward-pointing arrow indicates ICC < 0. The measurements calculated by comparing with a control condition are less reliable than by comparing with the implicit baseline in virtually every case.

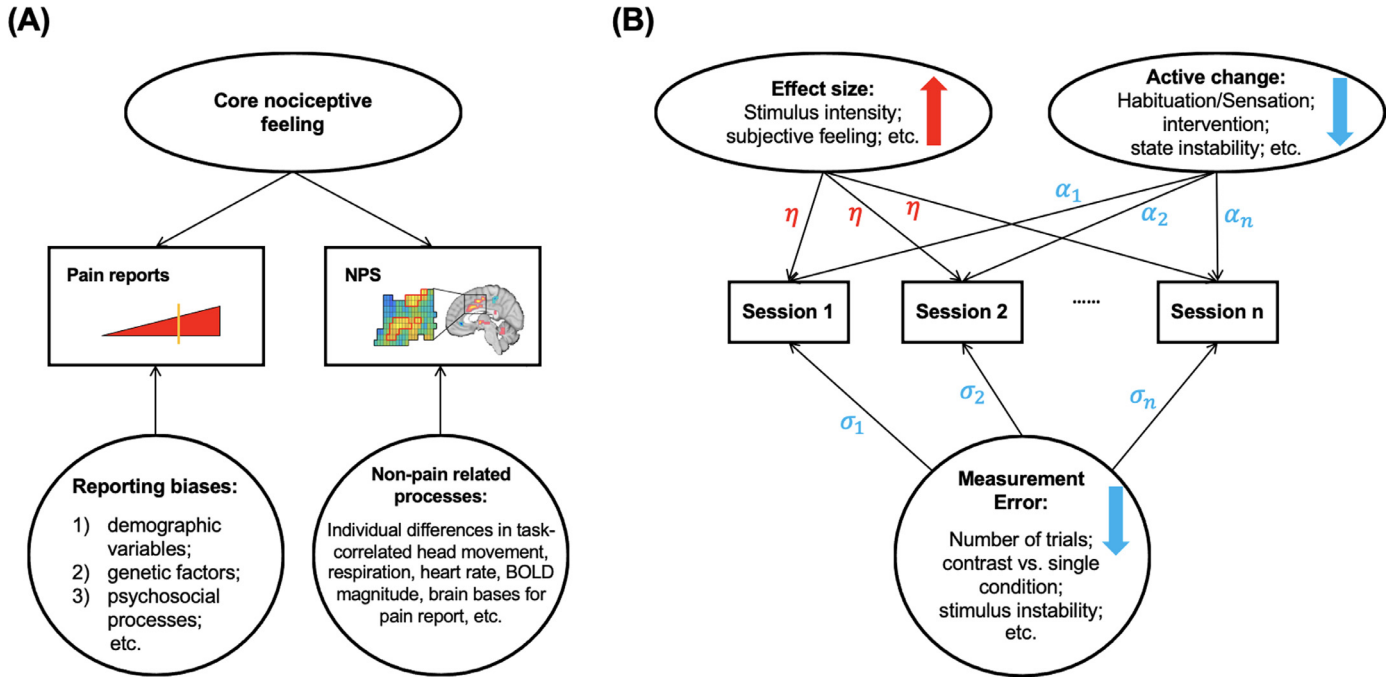


Fig. 3. Summary of variances and factors that influence the effect size and reliability. **(A) Different sources of variance at the between-person level for the NPS and self-report pain.** Rectangles represent the observed variables, i.e., pain reports and NPS. Ellipses represent the latent variables that we aim to measure, i.e., the core nociceptive feeling. The circle represents sources of variance that add to each observed measure. Both pain reports and NPS activity measure the core nociceptive circuits that generate pain experience. However, different sources of variance at the between-person level reduce the correlation between pain reports and the NPS response. This suggests that the NPS is not as useful as a surrogate measure for pain reports at the individual differences level. In contrast, the NPS could be useful as an objective biological target to measure physiological contributors to pain, in combination with subjective pain reports. **(B) Factors that influence reliability.** Rectangles represent the observed variables, such as the NPS response, across different sessions. Ellipses represent the latent variables that we are interested in modeling. Results suggest that stimuli with larger effect sizes have higher test-retest reliability, indicated by the upward-pointing red arrow, and have the same effect on all sessions, indicated by η. Some active change across sessions could decrease the test-retest reliability, indicated by the downward-pointing blue arrow. They might have different effects on different sessions, indicated by α_1, α_2, and α_n. The circle represents the measurement error that could decrease the test-retest reliability, indicated by the downward-pointing blue arrow. There might be different errors on different sessions, indicated by σ_1, σ_2, and σ_n.

could be led by the different sources of variance underlying the NPS responses and pain reports (see Fig. 3(A)). At the within-person level, different temperatures across trials are among the primary sources of variance in NPS responses and pain reports. The effect sizes of within-person correlations between the NPS and the temperatures were distributed from medium to huge (d = 0.53 - 2.67). The effect sizes of within-person correlations between the pain reports and the temperatures were distributed from very large to huge (d = 1.58 - 12.41; see Table S6). Both the NPS and pain reports are responsive to noxious stimuli intensities.

However, at the between-person level, the NPS and pain reports' variances may have been driven by many factors that are irrelevant to the stimuli intensities. One person can report more pain than another because of differences in demographic variables, genetic factors, and psychosocial processes (Fillingim, 2017; Woo and Wager, 2016). For example, individual differences in 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), respiration (Chang and Glover, 2009; Power et al., 2019), heart rate (Chang et al., 2009), BOLD magnitude (Levin et al., 2001) and inter-individual variation in brain bases for pain reports (Reddan and Wager, 2018). The combination of strong within-person correlations and only modest between-person correlations between the NPS and pain reports indicates that the NPS is not a surrogate for individual differences in pain reports. Instead, the NPS as an objective biological target could be useful for measuring pain in combination with subjective pain reports. For example, consider a clinical trial testing a new drug of analgesic, the investigators might want to know how the drug works independent of placebo effects, and possibly demonstrate brain penetrance and efficacy of the drug in affecting nociceptive pain-related systems (Duff et al., 2015). The NPS could be used to confirm these properties and may be particularly useful for doing so when self-reports are suspected to be influenced by placebo effects on systems (e.g., decision-making systems) other than those targeted by the drug (Tuttle et al., 2015; Zunhammer et al., 2018).

The effect sizes of between-person correlations should be interpreted with caution when estimated with a typical-sized neuroimaging sample (median N = 25). An advantage of assessing correlations across samples is that it provides a better estimate of the average correlation across studies, and its statistical significance, than small (and under-powered) individual studies. We did find a significant, moderately sized correlation between NPS and pain reports across studies (d = 0.49, r = 0.20). Only Study 4, with the smallest sample size (N = 17), showed a significant between-person correlation on its own (r = 0.74). However, we note that we cannot tell definitively whether some studies have significantly higher correlations than others, and the average effect size for between-person correlations (d = 0.49) is the best current estimate across studies.

One important question is whether the high correlation found in Study 4 is an overestimate due to the small sample size. Marek et al. (2020) found that estimating between-person correlations with small sample sizes leads to large sampling variability. Selection biases operate on this variability at both the region (feature) and study

(publication) level to inflate post hoc correlations estimates in reported effects (which are necessarily the largest; this is also illustrated in Reddan et al., 2017 JAMA Psychiatry). More than 2000 participants were recommended for stable between-person correlation estimation, though such sample sizes are difficult to obtain in novel exploratory studies. In the current paper, the sample sizes in individual studies were not large enough ($N <= 120$) provide a stable estimation of the between-person correlations between the NPS and pain reports (i.e., sampling variability is substantial at this sample size). However, the present study does not suffer from the selection bias issues that would lead to inflated post hoc correlations for two reasons: (1) we tested a single *a priori* measure (the NPS), eliminating selection bias at the feature-selection (e.g., region-selection) level, and (2) we report correlations in all datasets tested, precluding a study-level selection bias. Also, importantly, the studies in our sample were not selectively published based on between-participant correlations (which were never the principal focus of the original papers).

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. The higher reliabilities of the pain reports than the NPS responses were not due to larger inter-individual variances of the pain reports than the NPS responses (see Supplementary Result S1 and Figure S4). Test-retest reliability of pain reports has been extensively examined in previous pain-related studies that showed similar ICC values range from 0.75 - 0.96 (Jackson et al., 2020; Letzen et al., 2014, 2016; Upadhyay et al., 2015). Previous studies have examined the test-retest reliability of univariate brain responses 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), significantly activated clusters (0.33 - 0.74; Jackson et al., 2020) and functional connectivities (−0.17 - 0.77; Letzen et al., 2016). Compared with the previous univariate brain measures of pain, the NPS showed consistently high performance of short-term test-retest reliability across eight studies. It is noteworthy that although the short-term test-retest reliability is mathematically identical to the internal consistency reliability, they are conceptually different. Internal consistency measures how consistently a set of items, e.g., voxels in NPS, measures a particular construct, e.g., pain (Drost, 2011). At the same time, the short-term test-retest reliability characterizes the short-term temporal stability of measurement, e.g., the NPS response measured within a session (Drost, 2011). High values of internal consistency are not always desirable and could point to the redundancy of items (Streiner, 2003), while high test-retest reliability values are a desirable feature given that the constructs being measured are stable.

To test whether the NPS measure was stable across longer time scales, we examined two studies with 5-day and one-month intervals between sessions. We found that the NPS had high performance in longer-term test-retest reliability when evaluated with sufficient data per person. In our estimation, more than 60 trials per condition were required on average to achieve excellent test-retest reliability, though this was rarely done in practice (Chen et al., 2021; Dang et al., 2020; Rouder and Haaf, 2019). Recent studies also showed that to improve the reliability of the traditional univariate measures, having sufficient trials per person is more important than having a large sample size (Nee, 2019; Turner et al., 2018). However, when estimating the reliability of the univariate analyses (e.g., voxel-level), researchers usually use the thresholded brain maps and define the replication with some arbitrary standards (e.g., more than half of voxels in the cluster survived; Nee, 2019). The threshold to correct the multiple comparisons reduces power dramatically (Woo et al., 2014). Furthermore, the difficulty of defining a replication effect also brings uncertainty to the reliability estimation. By contrast, when estimating the reliability of the multivariate pattern signatures, such as the NPS, we computed a single scalar value for each brain map, avoiding the multiple comparison correction and improving the power. The multivariate pattern signatures can also specify a precise set of voxels and the topography of the relative expected activity levels across voxels, providing a basis for exact testing reliability.

Besides the number of trials per condition, we also found that reliability was improved with higher stimulus intensity (e.g., temperature) and when computing the [Pain > Baseline] contrast, rather than [High Pain > Low Pain] contrast (see Fig. 2 and Fig. 3(B)). In both these cases, a stronger fMRI contrast is present, leading to greater reliability. Additionally, the [Pain > Baseline] contrast includes non-specific brain responses to salient somatosensory stimuli, which likely also enhances reliability. Compared with high-temperature stimuli, low-temperature stimuli served as a control condition that activates non-nociceptive somatosensory pathways. The significant drop in reliability for [High vs. Low intensity] may suggest that part of what drives the NPS response when compared with rest is non-nociceptive somatosensory processes. However, the NPS has been found to have little temperature-related variability in the non-painful range (Wager et al., 2013). Also, reliability of NPS responses to low-temperature stimuli was reduced, indicating a stable nociceptive contribution. Another possibility is that the subtraction of a largely irrelevant, noisy variable (individual responses to low-temperature stimulation) adds error variance that reduces reliability. The (error) variance of the difference between high- and low-temperature responses is the sum of variances of the two conditions; thus, subtracting a variable measured with error is expected to increase the error variability in the difference score and decrease reliability.

In contrast to the NPS reliability, the reliability of the pain reports was less influenced by the trial number per condition, stimuli intensities, and contrast types (see Fig. 2). On average, the reliability of the pain reports achieved an excellent level with even one trial. While the reliability of NPS responses to largely non-painful low-temperature stimuli was reduced, pain reports were very reliable across all temperatures. A key insight is that nociceptive and non-nociceptive touch signals are separable, carried by largely distinct populations of neurons, so that they can be selectively impaired. There is much less signal in the NPS at non-painful stimulus intensities, and the relationship between stimulus intensity and NPS scores is weaker with non-painful stimulation (Wager et al., 2013). Thus, the NPS is expected to both respond and be reliable at painful stimulus intensities, but not necessarily non-painful intensities. In contrast, pain reports may be driven by both nociceptive and non-nociceptive somatosensory processes in a reliable fashion. Since pain reports are reliable at both high- and low-temperature stimuli, it is not surprising to find that the contrast types had less influence on the reliability of pain reports. The systematic differences between the NPS and pain reports indicated that they reflect different mixtures of underlying processes, further supporting the conclusion that the NPS is not simply a surrogate for pain reports.

The complete NPS performance was better than constituent local brain regions for both effect size and test-retest reliability. This finding is consistent with the argument that pain is encoded in distributed brain networks instead of a specific and isolated brain region (Petre et al., 2020; Woo and Wager, 2016). Interestingly, the six regions (i.e., bilateral insula, right dorsal posterior insula, aMCC, right S2, and right thalamus) with relatively larger effect sizes and reliabilities were the likely targets of ascending nociceptive afferents and activated in response to pain stimuli. Other local regions deactivated with pain and are not direct targets of nociceptive afferents have smaller effect sizes and reliabilities (Roy et al., 2012, 2014). The reliabilities of multivariate patterns of ROIs were heterogeneous (i.e., ICCs range from poor to excellent), similar to previous findings of pain-related ROIs using the univariate analyses (Letzen et al., 2014; Quiton et al., 2014; Upadhyay et al., 2015). In contrast, the reliabilities of the complete NPS were more homogeneous and all ranged from good to excellent level across multiple diverse studies.

The current study tests a large number of studies that are diverse in several aspects. Firstly, most of the studies contain some cognitive manipulations along with the painful stimuli, such as cognitive self-regulation intervention to increase or decrease pain (Woo et al., 2015), the combination of painful stimuli with visual or auditory cues for differ-

ent pain intensities (Atlas et al., 2010; Chang et al., 2015; Jepma et al., 2018; Roy et al., 2014), placebo manipulation (Roy et al., 2014). Secondly, the pain stimuli were applied to different body positions, including arm, foot, leg, and thumbnail, which were supposed to have different sensitivity to pain (Alburquerque-Sendín et al., 2018). Thirdly, the intensities of pain stimuli were largely varied regarding the temperature (44.3 - 50 ˚C) and duration (1.85 - 12.5 s). Lastly, the preprocessing pipelines and general linear models were diverse and maintained the same with the origin studies. This will likely reflect the variations of data analyses in the literatures. The diversities of the studies further support the generalizability of our findings about the measurement properties of the NPS and pain reports.

The participants in these studies were mainly young and healthy participants, with only one study testing participants with chronic back pain (i.e., study 10). The reliabilities of the NPS and pain reports in Study 10 were lower overall compared with Studies 1–9. The reliabilities of the NPS in the psychotherapy and placebo groups were lower than the control group numerically, though they were not significantly different from each other (see Table S3). The psychotherapy and placebo treatment targeted chronic back pain, while the fMRI task evoked thumb pain. The chronic back pain was more associated with activity in the affective and motivational systems tied to avoidance and less closely tied to systems encoding nociceptive input, such as the NPS (Ashar et al., 2021). Psychotherapy and placebo treatments may change brain processing of the affective and motivational systems of pain, and these changes introduce another source of variance, leading to lower test-retest reliability. We need further research to test the measurement properties of the NPS and subjective pain reports across more diverse participant samples, including clinical populations (Herr et al., 2011; Voepel-Lewis et al., 2010; Walton et al., 2011).

This paper focuses on the NPS because it has been one of the most extensively studied brain signatures for its validity and specificity in the pain domain (Chang et al., 2015; Krishnan et al., 2016; Ma et al., 2016; Van Oudenhove et al., 2020; Wager et al., 2013). Previous studies showed potential external validity of the NPS in clinical applications (Wager et al., 2013; López-Solà et al., 2017; McDermott et al., 2019; Weiber et al., 2019). Our findings suggested that trial numbers, contrast types, and stimuli intensities should be considered when designing the NPS measurement in clinical applications. There are still several challenges to extend the current findings to clinical measurement. (1) The NPS responses in the current paper were elicited by experimentally evoked pain, which might differ from clinical pain experience. The performance of the effect sizes and reliabilities of the NPS could be influenced by different types or aspects of clinical pain and need further research (Weiber et al., 2019). (2) The NPS's measurement properties have been primarily examined in healthy participants. Brain features related to clinical pain may be different and more heterogeneous (Ashar et al., 2021; Hashmi et al., 2013; Kutch et al., 2017; López-Solà et al., 2017; Tu et al., 2019). The tests in the current study characterizing the reliability, within-person, and between-person variances related to pain reports could be applied to any neuromarker, including other pain-related patterns (Brown et al., 2011; Geuter et al., 2020; Kucyi and Davis, 2015; Kutch et al., 2017; López-Solà et al., 2017; Marquand et al., 2010; Woo et al., 2017). Some other pain signatures possibly could have better performance in measurement properties than the NPS in different clinical populations.

Our study shows that it is crucial to characterize individual differences across studies and contexts. The correlation with individual differences in pain reports may vary across different experimental instructions and populations. For example, in study 4, we had a selected university population pre-screened for reliable pain reports and pre-calibrated for stimuli intensities and ended up with a very large effect in the between-person level correlation between the NPS and pain reports. The pre-calibration procedure provides people with more experience using the rating scale and calibrating their scale usage as they get a sense of the dynamic range of the stimuli. With little or no calibration procedure, subjective reports could be biased by anxiety or novelty effects, and/or subject to an initial elevation bias (Shrout et al., 2018). For studies using fixed temperatures, a restricted temperature range could in principle reduce reliability. If all temperatures are the same across individuals, then stimulus intensity-related variance will not drive individual differences, only differences in endogenous sensitivity and state effects to which the NPS appears to be largely insensitive (including expectations and biases related to relative judgments compared with previous stimuli, which vary trial to trial). We do not have sufficient data to directly compare the effects of calibration and training procedures on brain-pain correlations, but this is an important topic for future studies.

In sum, we find that both the NPS and pain reports have excellent test-retest reliability in a large sample of participants with diverse study procedures. As a measure of individual differences in pain sensitivity, the NPS is only modestly related to pain reports, suggesting that the NPS is not as useful as a surrogate measure of pain report. In contrast, the NPS could serve as an objective biological target to measure physiological contributors to pain, in combination with subjective pain reports or as a biological target in its own right. In the future, other multivariate brain patterns will need to be tested before used as translational biomarkers. Our study provides a blueprint for future studies performing such measurement properties testing and suggests factors that could improve test-retest reliability in future research.

Credit author statement

T.D.W. and X.H. conceived the research program. T.D.W., X.H., Y.K.A, P.K., B.P., V.S. analyzed the data. T.D.W. and X.H. wrote the paper. T.D.W., X.H., Y.K.A, P.K., B.P., V.S., L.Y.A., L.J.C., M.J., L.K., E.A.R.L., M.R., C.W. reviewed and edited the paper. T.D.W. supervised the entire work.

Data and code availability statements

Code for all analyses and figures is available at https://github.com/XiaochunHan/NPS_measurement_properties. Data for all analyses and figures is available at https://osf.io/v9px7/.

Acknowledgements

This project was supported by grants R01MH076136 (T.D.W.), R01DA046064, R01EB026549, and R01DA035484. Elizabeth A. Reynolds Losin was supported by a Mentored Research Scientist Development award from National Institute On Drug Abuse of the National Institutes of Health (K01DA045735). Lauren Y. Atlas was supported in part by funding from the Intramural Research Program of the National Center for Complementary and Integrative Health. Yoni K. Ashar was supported by NCATS Grant # TL1-TR-002386. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118844.

References

Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. Neuroimage 147, 736–745. doi:10.1016/j.neuroimage.2016.10.045.

Alburquerque-Sendín, F., Madeleine, P., Fernández-de-Las-Peñas, C., Camargo, P.R., Salvini, T.F., 2018. Spotlight on topographical pressure pain sensitivity maps: a review. J. Pain Res. 11, 215. doi:10.2147/JPR.S135769.

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. doi:10.1016/j.neuroimage.2016.02.079.

Ashar, Y.K., Gordon, A., Schubiner, H., Uipi, C., Knight, K., Anderson, Z., Carlisle, J., Polisky, L., Geuter, S., Flood, T.F., Kragel, P.A., Dimidjian, S., Lumley, M.A., Wager, T.D., 2021. Effect of pain reprocessing therapy vs placebo and usual care for patients with chronic back pain: a randomized clinical trial. JAMA Psychiatry doi:10.1001/jamapsychiatry.2021.2669.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851. doi:10.1016/j.neuroimage.2005.02.018.

Atlas, L.Y., Bolger, N., Lindquist, M.A., Wager, T.D., 2010. Brain mediators of predictive cue effects on perceived pain. J. Neurosci. 30, 12964–12977. doi: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, 1632–1648. doi:10.1016/j.pain.2014.05.015.

Bakdash, J.Z., Marusich, L.R., 2017. Repeated measures correlation. Front Psychol. 8, 456. doi:10.3389/fpsyg.2017.00456.

Barnhart, H.X., Haber, M.J., Lin, L.I., 2007. An overview on assessing agreement with continuous measurements. J. Biopharm. Stat. 17, 529–569. doi:10.1080/10543400701376480.

Bartoshuk, L.M., Duffy, V.B., Green, B.G., Hoffman, H.J., Ko, C.W., Lucchina, L.A., Marks, L.E., Snyder, D.J., Weiffenbach, J.M., 2004. Valid across-group comparisons with labeled scales: the gLMS versus magnitude matching. Physiol. Behav. 82, 109–114. doi:10.1016/j.physbeh.2004.02.033.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191, 133–155. doi:10.1111/j.1749-6632.2010.05446.x.

Bennett, C.M., Miller, M.B., 2013. fMRI reliability: influences of task and experimental design. Cogn. Affect. Behav. Neurosci. 13, 690–702. doi:10.3758/s13415-013-0195-1.

Brown, J.E., Chatterjee, N., Younger, J., Mackey, S., 2011. Towards a physiology-based measure of pain: patterns of human brain activity distinguish painful from non-painful thermal stimulation. PLoS ONE 6, e24124. doi:10.1371/journal.pone.0024124.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376. doi:10.1038/nrn3475.

Chang, C., Cunningham, J.P., Glover, G.H., 2009. Influence of heart rate on the BOLD signal: the cardiac response function. Neuroimage 44, 857–869. doi:10.1016/j.neuroimage.2008.09.029.

Chang, C., Glover, G.H., 2009. Relationship between respiration, end-tidal CO2, and BOLD signals in resting-state fMRI. Neuroimage 47, 1381–1393. doi:10.1016/j.neuroimage.2009.04.048.

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 Biol 13, e1002180. doi:10.1371/journal.pbio.1002180.

Chen, G., Padmala, S., Chen, Y., Taylor, P.A., Cox, R.W., Pessoa, L., 2021. To pool or not to pool: can we ignore cross-trial variability in FMRI? Neuroimage 225, 117496. doi:10.1016/j.neuroimage.2020.117496.

Cicchetti, D.V., Sparrow, S.A., 1981. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am. J. Ment. Def. 86, 127–137.

Cohen, J., 2013. Statistical Power Analysis For the Behavioral Sciences. Academic Press.

Craig, A.D., 2006. Retrograde analyses of spinothalamic projections in the macaque monkey: input to ventral posterior nuclei. J. Comp. Neurol. 499, 965–978. doi:10.1002/cne.21154.

Dang, J., King, K.M., Inzlicht, M., 2020. Why are self-report and behavioral measures weakly correlated? Trends Cogn. Sci. 24, 267–269. doi:10.1016/j.tics.2020.01.007.

Doyle, O.M., Mehta, M.A., Brammer, M.J., 2015. The role of machine learning in neuroimaging for drug discovery and development. Psychopharmacology (Berl.) 232, 4179–4189. doi:10.1007/s00213-015-3968-0.

Drost, E.A., 2011. Validity and reliability in social science research. Educ. Res. Perspect. 38, 105–123. https://search.informit.org/doi/10.3316/informit.491551710186460.

Dubois, J., Adolphs, R., 2016. Building a science of individual differences from fMRI. Trends Cogn. Sci. 20, 425–443. doi:10.1016/j.tics.2016.03.014.

Duff, E.P., Vennart, W., Wise, R.G., Howard, M.A., Harris, R.E., Lee, M., Wartolowska, K., Wanigasekera, V., Wilson, F.J., Whitlock, M., Tracey, I., Woolrich, M.W., Smith, S.M., 2015. Learning to identify CNS drug action and efficacy using multistudy fMRI data. Sci. Transl. Med. 7, 1–18. 10.1126/scitranslmed.3008438.

Elliott, M.L., Knodt, A.R., Cooke, M., Kim, M.J., Melzer, T.R., Keenan, R., Ireland, D., Ramrakha, S., Poulton, R., Caspi, A., Moffitt, T.E., Hariri, A.R., 2019. General functional connectivity: shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. Neuroimage 189, 516–532. doi:10.1016/j.neuroimage.2019.01.068.

Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Sison, M.L., Moffitt, T.E., Caspi, A., Hariri, A.R., 2020. What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychol. Sci. 31, 792–806. doi:10.1177/0956797620916786.

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. Dev. Cogn. Neurosci. 25, 58–68. doi:10.1016/j.dcn.2017.01.011.

Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S.S., Wright, J., Durnez, J., Poldrack, R.A., Gorgolewski, K.J., 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods. 16, 111–116. doi:10.1038/s41592-018-0235-4.

FDA-NIH Biomarker Working Group, 2016. BEST (Biomarkers, EndpointS, and Other Tools) Resource. Food and Drug Administration (US).

Fillingim, R.B., 2017. Individual differences in pain: understanding the mosaic that makes pain personal. Pain 158, S11. doi:10.1097/j.pain.0000000000000775.

Gabrieli, J.D.E., Ghosh, S.S., Whitfield-Gabrieli, S., 2015. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron 85, 11–26. doi:10.1016/j.neuron.2014.10.047.

Geuter, S., Reynolds Losin, E.A., Roy, M., Atlas, L.Y., Schmidt, L., Krishnan, A., Koban, L., Wager, T.D., Lindquist, M.A., 2020. Multiple brain networks mediating stimulus-pain relationships in humans. Cereb. Cortex. 30, 4204–4219. doi:10.1093/cercor/bhaa048.

Gordon, E.M., Laumann, T.O., Gilmore, A.W., Newbold, D.J., Greene, D.J., Berg, J.J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J.M., Coalson, R.S., Nguyen, A.L., McDermott, K.B., Shimony, J.S., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., Dosenbach, N.U.F., 2017. Precision functional mapping of individual human brains. Neuron 95, 791–807. doi:10.1016/j.neuron.2017.07.011.

Gratton, C., Kraus, B.T., Greene, D.J., Gordon, E.M., Laumann, T.O., Nelson, S.M., Dosenbach, N.U.F., Petersen, S.E., 2020. Defining individual-specific functional neuroanatomy for precision psychiatry. Biol. Psychiat. 88, 28–39. doi:10.1016/j.biopsych.2019.10.026.

Haynes, J.D., 2015. A primer on pattern-based approaches to fmri: principles, pitfalls, and perspectives. Neuron 87, 257–270. doi:10.1016/j.neuron.2015.05.025.

Harrison, O.K., Hayen, A., Wager, T.D., Pattinson, K.T., 2021. Investigating the specificity of the neurologic pain signature against breathlessness and finger opposition. Pain 1–12. doi:10.1097/j.pain.0000000000002327.

Hashmi, J.A., Baliki, M.N., Huang, L., Baria, A.T., Torbey, S., Hermann, K.M., Schnitzer, T.J., Apkarian, A.V., 2013. Shape shifting pain: chronification of back pain shifts brain representation from nociceptive to emotional circuits. Brain 136, 2751–2768. doi:10.1093/brain/awt211.

Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods. 50, 1166–1186. doi:10.3758/s13428-017-0935-1.

Herr, K., Coyne, P.J., McCaffery, M., Manworren, R., Merkel, S., 2011. Pain assessment in the patient unable to self-report: position statement with clinical practice recommendations. Pain Manag. Nurs. 12, 230–250. doi:10.1016/j.pmn.2011.10.002.

Herting, M.M., Gautam, P., Chen, Z., Mezher, A., Vetter, N.C., 2018. Test-retest reliability of longitudinal task-based fMRI: implications for developmental studies. Dev. Cogn. Neurosci. 33, 17–26. doi:10.1016/j.dcn.2017.07.001.

Jackson, J.B., O'Daly, O., Makovac, E., Medina, S., Rubio, A.L., McMahon, S.B., Williams, S.C.R., Howard, M.A., 2020. Noxious pressure stimulation demonstrates robust, reliable estimates of brain activity and self-reported pain. Neuroimage 221, 117178. doi:10.1016/j.neuroimage.2020.117178.

Jepma, M., Koban, L., van Doorn, J., Jones, M., Wager, T.D., 2018. Behavioural and neural evidence for self-reinforcing expectancy effects on pain. Nat. Hum. Behav. 2, 838–855. doi:10.1038/s41562-018-0455-8.

Kievit, R.A., Frankenhuis, W.E., Waldorp, L.J., Borsboom, D., 2013. Simpson's paradox in psychological science: a practical guide. Front. Psychol. 4, 513. doi:10.3389/fpsyg.2013.00513.

Koban, L., Jepma, M., López-Solà, M., Wager, T.D., 2019. Different brain networks mediate the effects of social and conditioned expectations on pain. Nat. Commun. 10, 4096. doi:10.1038/s41467-019-11934-y.

Koo, T.K., Li, M.Y., 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J. Chiropr. Med. 15, 155–163. doi:10.1016/j.jcm.2016.02.012.

Kraemer, H.C., 2014. The reliability of clinical diagnoses: state of the art. Annu. Rev. Clin. Psycho. 10, 111–130. doi:10.1146/annurev-clinpsy-032813-153739.

Kragel, P.A., Koban, L., Barrett, L.F., Wager, T.D., 2018. Representation, Pattern Information, and Brain Signatures: from Neurons to Neuroimaging. Neuron 99, 257–273. doi:10.1016/j.neuron.2018.06.009.

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). Psychol. Sci. 32, 622–626. doi:10.1177/0956797621989730.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12, 535–540. doi:10.1038/nn.2303.

Krishnan, A., Woo, C.W., Chang, L.J., Ruzic, L., Gu, X., López-Solà, M., Jackson, P.L., Pujol, J., Fan, J., Wager, T.D., 2016. Somatic and vicarious pain are represented by dissociable multivariate brain patterns. Elife. 5, e15166. doi:10.7554/eLife.15166.

Kucyi, A., Davis, K.D., 2015. The dynamic pain connectome. Trends Neurosci. 38, 86–95. doi:10.1016/j.tins.2014.11.006.

Kutch, J.J., Ichesco, E., Hampson, J.P., Labus, J.S., Farmer, M.A., Martucci, K.T., Ness, T.J., Deutsch, G., Apkarian, A.V., Mackey, S.C., Klumpp, D.J., Schaeffer, A.J., Rodriguez, L.V., Kreder, K.J., Buchwald, D., Andriole, G.L., Lai, H.H., Mullins, C., Kusek, J.W., Landis, J.R., Mayer, E.A., Clemens, J.Q., Clauw, D.J., Harris, R.E., 2017. Brain signature and functional impact of centralized pain: a multidisciplinary approach to the study of chronic pelvic pain (MAPP) network study. Pain 158, 1979. 10.1097/j.pain.0000000000001001.

Letzen, J.E., Boissoneault, J., Sevel, L.S., Robinson, M.E., 2016. Test-retest reliability of pain-related functional brain connectivity compared with pain self-report. Pain 157, 546–551. 10.1097/j.pain.0000000000000356.

Letzen, J.E., Sevel, L.S., Gay, C.W., O'Shea, A.M., Craggs, J.G., Price, D.D., Robinson, M.E., 2014. Test-retest reliability of pain-related brain activity in healthy controls undergoing experimental thermal pain. J. Pain. 15, 1008–1014. doi:10.1016/j.jpain.2014.06.011.

Levin, J.M., Frederick, B.B., Ross, M.H., Fox, J.F., von Rosenberg, H.L., Kaufman, M.J., Lange, N., Mendelson, J.H., Cohen, B.M., Renshaw, P.F., 2001. Influence of baseline hematocrit and hemodilution on BOLD fMRI activation. Magn. Reson. Imaging. 19, 1055–1062. doi:10.1016/S0730-725X(01)00460-X.

Lindquist, M.A., Krishnan, A., López-Solà, M., Jepma, M., Woo, C.W., Koban, L., Roy, M., Atlas, L.Y., Schmidt, L., Chang, L.J., Losin, E.A.R., Eisenbarth, H., Ashar, Y.K., Delk, E.,

Wager, T.D., 2017. Group-regularized individual prediction: theory and application to pain. Neuroimage 145, 274–287. doi:10.1016/j.neuroimage.2015.10.074.
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, S187–S198. doi:10.1016/j.neuroimage.2008.10.065.
Lohani, S., Poplawsky, A.J., Kim, S.G., Moghaddam, B., 2017. Unexpected global impact of VTA dopamine neuron activation as measured by opto-fMRI. Mol psychiatry 22, 585–594. doi:10.1038/mp.2016.102.
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, 34–47. 10.1097/j.pain.0000000000000707.
Losin, E.A.R., Woo, C.W., Medina, N.A., Andrews-Hanna, J.R., Eisenbarth, H., Wager, T.D., 2020. Neural and sociocultural mediators of ethnic differences in pain. Nat. Hum. Behav. 4, 517–530. doi:10.1038/s41562-020-0819-8.
Manuck, S.B., Brown, S.M., Forbes, E.E., Hariri, A.R., 2007. Temporal stability of individual differences in amygdala reactivity. Am. J. Psychiatry. 164, 1613–1614. doi:10.1176/appi.ajp.2007.07040609.
Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Feczko, E., Miranda-Dominguez, O., Graham, A.M., Earl, E.A., Perrone, A.J., Cordova, M., Doyle, O., Moore, L.A., Conan, G., Uriarte, J., Snider, K., Tam, A., Chen, J., Newbold, D.J., Zheng, A., Seider, N.A., Van, A.N., Laumann, T.O., Thompson, W.K., Greene, D.J., Petersen, S.E., Nichols, T.E., Yeo, B.T.T., Barch, D.M., Garavan, H., Luna, B., Fair, D.A., Dosenbach, N.U., 2020. Towards reproducible brain-wide association studies. BioRxiv doi:10.1101/2020.08.21.257758.
Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., Mourão-Miranda, J., 2010. Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. Neuroimage 49, 2178–2189. doi:10.1016/j.neuroimage.2009.10.072.
Ma, Y., Wang, C., Luo, S., Li, B., Wager, T.D., Zhang, W., Rao, Y., Han, S., 2016. Serotonin transporter polymorphism alters citalopram effects on human pain responses to physical pain. Neuroimage 135, 186–196. doi:10.1016/j.neuroimage.2016.04.064.
Mazzola, L., Isnard, J., Peyron, R., Mauguière, F., 2012. Stimulation of the human cortex and the experience of pain: wilder Penfield's observations revisited. Brain 135, 631–640. doi:10.1093/brain/awr265.
McDermott, L.A., Weir, G.A., Themistocleous, A.C., Segerdahl, A.R., Blesneac, I., Baskozos, G., Clark, A.J., Millar, V., Peck, L.J., Ebner, D., Tracey, I., Serra, J., Bennett, D.L., 2019. Defining the functional role of NaV1. 7 in human nociception. Neuron 101, 905–919. doi:10.1016/j.neuron.2019.01.047.
McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods. 1, 30–46. doi:10.1037/1082-989X.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, 2636–2643. doi:10.1016/j.neuroimage.2011.08.076.
Nakagawa, S., Schielzeth, H., 2010. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biol. Rev. 85, 935–956. doi:10.1111/j.1469-185X.2010.00141.x.
Nee, D.E., 2019. fMRI replicability depends upon sufficient individual-level data. Commun. Biol. 2, 1–4. doi:10.1038/s42003-019-0378-6.
Nichols, T.E., Das, S., Eickhoff, S.B., Evans, A.C., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M.P., Poldrack, R.A., Poline, J.B., Proal, E., Thirion, B., Essen, D.C.V., White, T., Yeo, B.T.T., 2017. Best practices in data analysis and sharing in neuroimaging using MRI. Nat. Neurosci. 20, 299–303. doi:10.1038/nn.4500.
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. doi:10.1016/j.neuroimage.2019.116157.
Nord, C.L., Gray, A., Charpentier, C.J., Robinson, O.J., Roiser, J.P., 2017. Unreliability of putative fMRI biomarkers during emotional face processing. Neuroimage 156, 119–127. doi:10.1016/j.neuroimage.2017.05.024.
O'Connor, D., Potler, N.V., Kovacs, M., Xu, T., Ai, L., Pellman, J., Vanderwal, T., Parra, L.C., Cohen, S., Ghosh, S., Escalera, J., Grant-Villegas, N., Osman, Y., Bui, A., Craddock, R.C., Milham, M.P., 2017. The Healthy Brain Network Serial Scanning Initiative: a resource for evaluating inter-individual differences and their reliabilities across scan conditions and sessions. Gigascience 6, 1–14. doi:10.1093/gigascience/giw011.
Orrù, 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. Neurosci. Biobehav. R. 36, 1140–1152. doi:10.1016/j.neubiorev.2012.01.004.
Pannunzi, M., Hindriks, R., Bettinardi, R.G., Wenger, E., Lisofsky, N., Martensson, J., Butler, O., Filevich, E., Becker, M., Lochstet, M., Kühn, S., Deco, G., 2017. Resting-state fMRI correlations: from link-wise unreliability to whole brain stability. Neuroimage 157, 250–262. doi:10.1016/j.neuroimage.2017.06.006.
Petre, B., Kragel, P.A., Atlas, L.Y., Geuter, S., Jepma, M., Koban, L., Krishnan, A., López-Solà, M., Roy, M., Woo, C.W., Wager, T.D., 2020. Evoked pain intensity representation is distributed across brain systems: a multistudy mega-analysis. BioRxiv doi:10.1101/2020.07.04.182873.
Plichta, M.M., Schwarz, A.J., Grimm, O., Morgen, K., Mier, D., Haddad, L., Gerdes, A.B.M., Sauer, C., Tost, H., Esslinger, C., Colman, P., Wilson, F., Kirsch, P., Meyer-Lindenberg, A., 2012. Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery. Neuroimage 60, 1746–1758. doi:10.1016/j.neuroimage.2012.01.129.
Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.B., Vul, E., Yarkoni, T., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. rev. neurosci. 18, 115. doi:10.1038/nrn.2016.167.
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. doi:10.1016/j.neuroimage.2019.116041.
Quiton, R.L., Keaser, M.L., Zhuo, J., Gullapalli, R.P., Greenspan, J.D., 2014. Intersession reliability of fMRI activation for heat pain and motor tasks. Neuroimage Clin 5, 309–321. doi:10.1016/j.nicl.2014.07.005.
Reddan, M.C., Lindquist, M.A., Wager, T.D., 2017. Effect Size Estimation in Neuroimaging. JAMA Psychiat 74, 207–208. doi:10.1001/jamapsychiatry.2016.3356.
Reddan, M.C., Wager, T.D., 2018. Modeling Pain Using fMRI: from Regions to Biomarkers. Neurosci. Bull. 34, 208–215. doi:10.1007/s12264-017-0150-1.
Rissman, J., Gazzaley, A., D'Esposito, M., 2004. Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage 23, 752–763. doi:10.1016/j.neuroimage.2004.06.035.
Rouder, J.N., Haaf, J.M., 2019. A psychometrics of individual differences in experimental tasks. Psychon. B. Rev. 26, 452–467. doi:10.3758/s13423-018-1558-y.
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. Nat. Neurosci. 17, 1607–1612. doi:10.1038/nn.3832.
Roy, M., Shohamy, D., Wager, T.D., 2012. Ventromedial prefrontal-subcortical systems and the generation of affective meaning. Trends Cogn. Sci. 16, 147–156. doi:10.1016/j.tics.2012.01.005.
Sawilowsky, S.S., 2009. New effect size rules of thumb. J. Mod. Appl. Stat. Methods. 8, 26. doi:10.22237/jmasm/1257035100.
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, 1957–1964. doi:10.1016/j.neuroimage.2009.08.040.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428. doi:10.1037/0033-2909.86.2.420.
Shrout, P.E., Stadler, G., Lane, S.P., McClure, M.J., Jackson, G.L., Clavél, F.D., Lida, M., Gleason, M.E.J., Xu, J.H., Bolger, N., 2018. Initial elevation bias in subjective reports. Proc. Natl. Acad. Sci. U.S.A. 115, E15–E23. doi:10.1073/pnas.1712277115.
Streiner, D.L., 2003. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J. Pers. Assess. 80, 99–103. doi:10.1207/S15327752JPA8001_18.
Tu, Y., Jung, M., Gollub, R.L., Napadow, V., Gerber, J., Ortiz, A., Lang, C., Mawla, I., Shen, W., Chan, S.T., Wasan, A.D., Edwards, R.R., Kaptchuk, T.J., Rosen, B., Kong, J., 2019. Abnormal medial prefrontal cortex functional connectivity and its association with clinical symptoms in chronic low back pain. Pain 160, 1308–1318. 10.1097/j.pain.0000000000001507.
Turner, B.O., Paul, E.J., Miller, M.B., Barbey, A.K., 2018. Small sample sizes reduce the replicability of task-based fMRI studies. Commun. Biol. 1, 1–10. doi:10.1038/s42003-018-0073-z.
Tuttle, A.H., Tohyama, S., Ramsay, T., Kimmelman, J., Schweinhardt, P., Bennett, G.J., Mogil, J.S., 2015. Increasing placebo responses over time in U.S. clinical trials of neuropathic pain. Pain 156, 2616–2626. 10.1097/j.pain.0000000000000333.
Upadhyay, J., Lemme, J., Anderson, J., Bleakman, D., Large, T., Evelhoch, J.L., Hargreaves, R., Borsook, D., Becerra, L., 2015. Test-retest reliability of evoked heat stimulation BOLD fMRI. J. Neurosci. Meth. 253, 38–46. doi:10.1016/j.jneumeth.2015.06.001.
Van Oudenhove, L., Kragel, P.A., Dupont, P., Ly, H.G., Pazmany, E., Enzlin, P., Rubio, A., Delon-Martin, C., Bonaz, B., Aziz, Q., Tack, J., Fukudo, S., Kano, M., Wager, T.D., 2020. Common and distinct neural representations of aversive somatic and visceral stimulation in healthy individuals. Nat. Commun. 11, 1–11. doi:10.1038/s41467-020-19688-8.
Voepel-Lewis, T., Zanotti, J., Dammeyer, J.A., Merkel, S., 2010. Reliability and validity of the face, legs, activity, cry, consolability behavioral tool in assessing acute pain in critically ill patients. Am. J. Crit. Care. 19, 55–61. doi:10.4037/ajcc2010624.
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 Engl. J. Med. 368, 1388–1397. doi:10.1056/NEJMoa1204471.
Wager, T.D., Nichols, T.E., 2003. Optimization of experimental design in fMRI: a general framework using a genetic algorithm. Neuroimage 18, 293–309. doi:10.1016/S1053-8119(02)00046-0.
Walton, D.M., Macdermid, J.C., Nielson, W., Teasell, R.W., Chiasson, M., Brown, L., 2011. Reliability, standard error, and minimum detectable change of clinical pressure pain threshold testing in people with and without acute neck pain. J. Orthop. Sports Phys. Ther. 41, 644–650. https://www.jospt.org/doi/10.2519/jospt.2011.3666.
Weber II, K.A., Wager, T.D., Mackey, S., Elliott, J.M., Liu, W.C., Sparks, C.L., 2019. Evidence for decreased Neurologic Pain Signature activation following thoracic spinal manipulation in healthy volunteers and participants with neck pain. NeuroImage Clin 24, 102042. doi:10.1016/j.nicl.2019.102042.
Woo, C.W., Chang, L.J., Lindquist, M.A., Wager, T.D., 2017a. Building better biomarkers: brain models in translational neuroimaging. Nat. Neurosci. 20, 365–377. doi:10.1038/nn.4478.
Woo, C.W., Krishnan, A., Wager, T.D., 2014. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. Neuroimage 91, 412–419. doi:10.1016/j.neuroimage.2013.12.058.
Woo, C.W., Roy, M., Buhle, J.T., Wager, T.D., 2015. Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. PLoS Biol 13, e1002036. doi:10.1371/journal.pbio.1002036.
Woo, C.W., Schmidt, L., Krishnan, A., Jepma, M., Roy, M., Lindquist, M.A., Atlas, L.Y., Wager, T.D., 2017b. Quantifying cerebral contributions to pain beyond nociception. Nat. Commun. 8, 14211. doi:10.1038/ncomms14211.
Woo, C.W., Wager, T.D., 2016. What reliability can and cannot tell us about pain report and pain neuroimaging. Pain 157, 511–513. doi:10.1097/j.pain.0000000000000442.

Xu, T., Opitz, A., Craddock, R.C., Wright, M.J., Zuo, X.N., Milham, M.P., 2016. Assessing variations in areal organization for the intrinsic brain: from fingerprints to reliability. Cereb. Cortex. 26, 4192–4211. doi:10.1093/cercor/bhw241.

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. doi:10.1016/j.neuroimage.2019.04.060.

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 Neurol. 75, 1321–1330. doi:10.1001/jamaneurol.2018.2017.

Zuo, X.N., Xing, X.X., 2014. Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective. Neurosci. Biobehav. R. 45, 100–118. doi:10.1016/j.neubiorev.2014.05.009.

Zuo, X.N., Xu, T., Milham, M.P., 2019. Harnessing reliability for neuroscience research. Nat. Hum. Behav. 3, 768–771. doi:10.1038/s41562-019-0655-x.