**Discussion** (count: 1018 words)

Effect size difference between univariate and multivariate brain pattern

Start here:

Current efforts towards translation… need large effects and reliability...recently renewed interests… effect sizes important because… Reliability is important because…(two sentences)

We studied reliability across…. Xx studies and xxx patients. Focused on NPS (a measure of…) because…

We characterized effect size and test-retest reliability. - (1) Effect sizes Within-person and between, for predicting pain reports and stimulus intensity

(2) Reliability of multivariate pattern response vs. individual sub-regions. (3) Factors that influence reliability, including within-person measurement error, amount of data per individual, test-retest interval, effect size in engaging the target measure (e.g., temperature, as NPS responds at high temperature), contrast with baseline vs. an active control. We also assessed NPS reliability across days to weeks in both healthy sample and chronic pain sample. WE Assessed and compared reliability of NPS and pain reports, which also allowed us to assess whether NPS and pain reports measure the same thing, or are different constructs.

The NPS, showed high performance in the test-retest reliability. The short-term (i.e., within one day) test-retest reliability of NPS which were tested in nine studies are distributed from good to excellent (\*\*\*numbers\*\*\*). Given the same number of trials, the longer time interval between sessions didn’t impair the test-retest reliability of NPS in another two studies with one-week and one-month interval between tests. The length of time interval between tests is critical for the test-retest reliability of a dynamic state measurement, as the state can vary across time (Streiner 2003). While for a trait measurement, the test-retest reliability is less sensitive to the time interval. For example, state anxiety has short-term reliability around 0.9, but longer-term (on the order of 1 month) test-retest reliability of 0.1-0.4 (Spielberger 1983). The high performance of NPS in the long-term test-retest reliability indicates that NPS is not just a measure of state, but a stable and reliable trait in the long term ( ).

Though the test-retest reliability for both NPS and subjective pain rating were high, the correlations between NPS and pain reports across participants were widely varied and not high in general. Subjective pain reports reflect a complex mix of brain and psychological processes. One person can report more pain than another because of differences in nociception, emotion, decision making, self-awareness, social cognition, and communicative tendencies (Woo and Tor, 2016). In contrast, the NPS relative purely reflects pain experience in response to noxious stimuli (Wager et al., 2013; Brascher et al., 2016; Krishnan et al., 2016; Krishnan et al., 2016; Ma et al., 2016). These two measurements might be measuring divergent mental processes in many studies. The individual difference of NPS response might reflect variance in pain experience, while the individual difference of subjective pain reports might reflect communicative bias, such as “stoics” vs “communicators”. It is not a problem for the within-subject correlation between NPS and subjective pani reports, and they are significant across all nine studies. Two studies with high between-subject correlations (i.e. exp and nsf) enrolled temperature calibration procedures in the experiment manipulations, which adjusted the pain level based on subjective pain reports, instead of using the objective temperatures. This procedure might help to diminish the divergence between two measures.

Variance between Individuals is one of the main factors that influences the test-retest reliability (i.e., ICC) calculation and another main factor is the measurement error variance. Given the consistent variance between individuals in our samples, we tested several factors that might contribute to the measurement error variance. Firstly, the measurement averaged with more trials is more reliable than with less trials. Trial-by-trial variance is one source of error variance and could be dramatically improved by increasing the number of trials. In our estimation, more than 60 trials per condition are required to achieve excellent test-retest reliability. However, trial numbers in most of the fMRI studies are limited considering the scanning expense and large head-motion accompanied by fatigue of participants ( ). Given the same number of trials in the experiment, the stimuli with larger effect size, i.e., higher temperature of the thermal stimuli, produce more reliable measurement. This might due to high intensity stimuli are more salient for participants and increase the signal-to-noise ratio ( ). Moreover, the measurement calculated in contrast with the baseline is more reliable than in contrast with a control condition, especially when the reliability of the control condition is low. All these factors have a larger influence on the NPS than self-reported pain, which supports the argument that fMRI measures are noisier than self-reported pain (Letzen et al., 2016, Woo and Tor, 2016).

Again, the test-retest reliability characterized by ICC values relates the inherent variability in the ‘true’ values between subjects to the magnitude of the measurement error in observed measurements. If reliability is high, measurement errors are small in comparison to the true differences between subjects, so that subjects can be relatively well distinguished based on the observed measurements. Conversely, if measurement errors are large compared with the true differences between subjects, differences between measurements of subjects could be purely led by error rather than by a genuine difference in their true values, thus the reliability is low in this situation. Therefore, multivariate brain biomarkers with high test-retest reliability are suitable for measuring individual differences of the corresponding mental processes or clinical symptoms, such as prognostic biomarkers for future disease and predictive biomarkers for treatment response (FDA-NIH Biomarker Working Group 2016). However, not all biomarkers are developed to measure individual difference. Some other biomarkers are mainly focused on detecting variation within individuals such as detecting states of pain, consciousness, disease, etc that varies across time. For those biomarkers, it is more appropriate to evaluate the within-subject measurement error instead of the test-retest reliability. It is necessary to examine different reliability measures according to the applying purpose of the biomarkers.

Reliability is not a fixed property of a measurement or a measurement technology, such as fMRI; rather it is a property of the scores on a measurement for a particular sample of participants (Wilkinson & The Task Force on Statistical Inference, 1999). Besides the measurement itself, reliability could be influenced by a number of attributes of the samples, such as the sample size and heterogeneity of samples. The more heterogeneous the sample, then the larger variance between subjects, and higher reliability of measurement (Henson, Kogan and Vacha-Haase, 2001). Therefore, it is important for researchers to assess reliability for their own data. The data in the current study is quite diverse in the aspects of the noxious stimuli (such as temperature, duration, body location, etc.), experiment manipulation (such as expectation, self-regulation, etc.), population (such a number of groups of healthy participants, and chronic pain patients), scanning settings (such as different scanners and parameters) and so on. We believe the reliability of NPS we get in the current study is quite representative. In the future, there are a lot more reliabilities of multivariate brain patterns that need to be tested before being used as biomarkers. Our study provides a

blueprint for future reliability testing and points out factors that could be considered to improve reliability.