What reliability can and cannot tell us about pain report and pain neuroimaging

Choong-Wan Woo[1,2*], Tor D. Wager[1,2]

[1]Department of Psychology and Neuroscience, University of Colorado, Boulder

[2]Institute of Cognitive Science, University of Colorado, Boulder

* Please address correspondence to:

Choong-Wan Woo

Department of Psychology and Neuroscience

University of Colorado Boulder

345 UCB, Boulder, CO 80305

Email: Choongwan.Woo@Colorado.Edu

Telephone: (720) 443-3640

In this issue of Pain, Letzen et al. [3] examine the test-retest reliability of fMRI connectivity (fcMRI) measures and self-reported pain during pain stimulation, and show that fcMRI measures are less reliable than self-reported pain. This study is a valuable endeavor, and as we move towards developing and using neuroimaging-based biomarkers for clinical purposes (e.g., making predictions and decisions about an individual), efforts to establish reliability and reproducibility of candidate biomarkers will become more and more critical.

Based on the results, Letzen et al. [3] concluded, and we agree, that fMRI measures are noisier than self-reported pain. This might be a big issue if we view fMRI measures as a substitute for pain ratings [6]. However, the real value of using brain measures to study pain is not in replacing pain ratings, but in serving to (a) provide a better understanding of how our mind and body generate and regulate pain, and (b) provide ways to see and measure its component neurobiological processes. The reason that we need brain markers for pain is not that pain ratings are "flawed" in their reliability or that they are "flawed" at all, but that pain ratings reflect a complex mix of brain and psychological processes. For example, one person can report more pain than other because of differences in nociception, emotion, decision-making, self-awareness, social cognition, and communicative tendencies. Because there is no single process that causes people to report more or less pain, self-reported pain provides only limited clues on what the underlying causes and what the best course of treatment might be. Many other disorders are similarly heterogeneous, and symptoms alone have not proven to be sufficient to guide effective treatment. In cancer, for example, diagnosis and treatment have progressively shifted from symptoms and overt signs to molecular subtypes that respond strongly to tailored molecular treatments [1].

Reliability is an important measurement property, though it is just one piece of the puzzle. Principally, reliability places constraints on the utility of a measure for assessing individual differences. However, those constraints are more subtle than it first appears. Below, we briefly elaborate on what reliability is and what constraints it does and does not place on the use of fMRI in assessment and personalized medicine.

Reliability: More is not always better

There are many different ways to define reliability, but the most common definition of reliability, which is also used in Letzen et al. [3], is based on the ratio of between-person to within-person variability across repeated measures [7]. Simply put, if a measure shows stable individual differences over time, the measure has high reliability. This is a specific type of reliability, and other types include repeatability (stability within a same subject over a short period of time) and reproducibility (stability across different days, labs, scanners, etc.). All of these characteristics are considered by the FDA in assessing the quality of measures [8].

Why is reliability only a piece of the puzzle then? The reliability assessed in Letzen et al. [3] concerns the stability of individual differences over time. Therefore, any factor that increases between-person variability, even if it is related to artifacts and biases, will increase reliability, whereas any factor that reduces between-person variability will reduce reliability, if the artifacts or biases persist or are otherwise stable across repeated assessments. This leads to a number of factors that potentially affect reliability in the ways that we may not want, revealing that "more is not always better."

For example, if a sample consists of a more heterogeneous population, measures will tend to show higher reliability on average. Such heterogeneity might include variability in standards for pain reporting across different cultures ('stoics' vs. 'communicators') [5] and other communicative biases (see **Figure 1**). Such differences will increase the reliability in pain ratings at the cost of biological meaningfulness and homogeneity of treatment responses. Even more troublingly, factors such as inconsistency in how experimenters communicate with participants and explain tasks and rating procedures could also *increase* the reliability of pain ratings if their effects persist across test and retest. Imagine the case where different experimenters are assigned to different participants, but the same experimenter does both test and retest for each participant. In this case, experimenter effects will increase between-person variability and thus the reliability of pain ratings. If participants are tested by different experimenters across assessments, this will reduce reliability on average. However, if early testing effects (e.g., a particularly competent and soothing experimenter) carry over to influence subsequent tests, then experimenter effects could increase reliability in this case as well. Conversely, study samples highly trained on how to use pain rating scales will, on average, show lower reliability than samples with poor training because good training will reduce between-person variability, whereas poor training will increase between-person variability.

In the context of fMRI, factors that have varied effects across participants (increasing between-person variability) but consistent effects within a subject (reducing within-person variability) will increase the reliability of fMRI measures. For example, misalignment of brains across individuals, individual differences in head movement, individual differences in physiological artifacts, individual differences in the shape of hemodynamic responses, and other undesirable sources of stable between-person variability will all *increase* reliability. Conversely, if there is a brain region that behaves very similarly (e.g., responds similarly to painful stimulation) in everyone, the reliability of the region is likely to be poor because of its low between-person variability. Therefore, in addition to seeking for measures with high reliability of individual differences, it is critical to understand what those individual differences mean and what constructs they measure.

What brain measures should we be assessing the reliability of? From regions to distributed patterns

Letzen et al. [3] correctly pointed out that fMRI-based markers are composed of selected features, and they assessed the reliability of some selected features, in particular functional correlations across a set of regions of interest (ROIs). In general, the reliability of features places constraints on the reliability of the markers as a whole. However, this does not occur in a linear fashion, and so Letzen et al.'s [3] analyses do not have much bearing on how reliable *optimized* brain markers for pain are likely to be. Below, we describe two principles that constrain how feature reliability determines marker reliability: *feature selection* and the *"wisdom of crowds"*.

Feature selection. Reliability varies across regions (or pairs of regions), but it is uncertain which regions are most relevant for pain. Reliability in connectivity among

Letzen et al.’s [3] regions had intra-class correlation coefficients (a measure of reliability) ranging from -0.17 (poor) to 0.77 (good). But which of these correlations are relevant for pain, and which are unrelated to pain? If the regions with poor reliability are unrelated to pain, they will be ignored by optimized predictive models (e.g., classifiers) anyway. Alternatively, if the poor-reliability regions are precisely the most useful ones for assessing pain, markers for pain will likely be unreliable as well. Further work is needed to answer this question.

However, there are reasons why reliability might be low with the ROI-based approach that is used in Letzen et al. [3], even if the reliability of optimized markers is high. Because they pre-defined multi-voxel regions, there is always a possibility that the regions averaged over heterogeneous voxels, an effect called ‘partial voluming.’ FcMRI in particular requires identifying the precise voxels that are most meaningfully connected to other target regions. With 100,000 voxels, there are 5 trillion (100,000 × (100,000-1) / 2) possible pairwise connectivity values in the brain. Some of these will be more reliable, others less so. Machine learning approaches can pick out relevant, and sometimes explicitly the most reliable (e.g., [2]), features from among this massive number of possibilities.

Wisdom of crowds. FMRI biomarker development often involves optimization of large-scale multivariate, predictive models across multiple brain regions. If a mental process is distributed across multiple brain regions, multivariate models of patterns distributed across those brain systems will perform best as markers. Pain, even in its most elementary, nociception-driven forms, is thought to be distributed in this fashion. In such cases, the average across a number of predictive regions will be asymptotically more reliable than any single component region or feature. To illustrate this, we tested the reliability of the Neurologic Pain Signature [9], a distributed marker involving a weighted average across multiple brain regions, and compared it to the reliability of pain-related regions of interest (see **Figure 2**). As expected, reliability is higher for NPS responses than single regions, and in this case was comparable to the reliability of self-report.

Finally, the performance of a brain marker in predicting individual differences in pain places a lower bound on reliability [4]: For example, a measure cannot asymptotically predict pain (or any other external outcome) with *r* = 0.7 if its reliability (correlation with itself across repeated measures) is less than 0.7. This suggests that there may be grounds for optimism when it comes to whole-brain pattern signatures considering their high correlations with pain (e.g., correlation between the NPS and pain ratings = 0.74 [9]).

Conclusion

Pain ratings are reliable under many circumstances, but reliability cannot tell us what pain ratings actually measure. In some cases, pain ratings can reflect cognitive biases, and in others, willingness to communicate or desire to be stoic. This is not a "flaw," but a feature of complex human behavior. The value of brain markers, whether reliable or not, is in measuring neurophysiological processes that are more closely and consistently aligned with particular 'ingredients' of pain, whether they be nociception, affect, or judgment and decision-making. Understanding which brain processes relate to which ingredients is a grand challenge, but one worth undertaking.

Conflict of interest statement

The authors have no conflicts of interest or financial relationships relevant to this commentary to disclose.

Acknowledgments

This work was funded by NIDA R01DA035484-01 (T.D.W, P.I.).

Figure Legend

Figure 1. An illustration of the effects of sample heterogeneity on test-retest reliability of pain ratings. Reliability is generally a desirable property for any measure. However, any stable individual difference can increase reliability, independent of the validity and utility of the measure in other respects. To illustrate this, we simulated intra-class correlation coefficients (ICCs), a measure of reliability, in samples (N = 30) consisting of one homogeneous population (green) or a mix of "communicators" that report higher pain and "stoics" that report less pain (orange and yellow, respectively). In the simulation, the mean pain ratings were set to 50, 55, and 45 for the homogeneous, "communicator," and "stoic" groups. The between-person and within-person (across test/re-test) standard deviations were set to 5 for all groups; thus, the noise levels were the same for both samples. We repeated the simulation 1000 times and observed the effects on the distribution of test-retest ICC values (right). The results show that more heterogeneous samples showed higher reliability (mean ICC = 0.80) than more homogeneous samples (mean ICC = 0.65). The principle of stable heterogeneity increasing reliability applies to both individual differences of interest (e.g., genetic differences in pain sensitivity) and nuisance variables (e.g., the use of multiple experimenters or variable testing procedures), as long as their effects carry over or are otherwise stable from test to retest. The simulation code is available at https://github.com/wanirepo/Woo_TRR_commentary_PAIN

Figure 2. Test-retest reliability of ROIs vs. whole-brain pattern signature for pain. (A) Using a previously published dataset (*N* = 33) [10], we calculated test-retest reliability (using intra-class correlation [ICC]) for a) self-reported pain; b) pattern expression values of the Neurologic Pain Signature (NPS), a pattern optimized to track the intensity of pain experience [9]; and c-e) averaged fMRI activity within three *a priori* regions-of-interest (ROIs). In this analysis, we included only trials with heat in the noxious range for all participants (47.3 and 48.3 °C). **(B)** Results showed that the ICC coefficients for pain ratings and NPS response are within the "good" range—and roughly comparable—whereas the ICC coefficients for ROI averages are "fair." This study included a relatively homogeneous group

of participants (which reduces reliability) by design, so reliability in more heterogeneous community and patient populations must be tested in future work. dACC, dorsal anterior cingulate cortex. L Ins, left insular cortex. R Ins, right insular cortex.

References

[1] Cloughesy TF, Cavenee WK, Mischel PS. Glioblastoma: From Molecular Pathology to Targeted Treatment. Annu. Rev. Pathol. Mech. Dis. 2014;9:1–25.

[2] Craddock RC, Holtzheimer PE, Hu XP, Mayberg HS. Disease state prediction from resting state functional connectivity. Magn. Reson. Med. Off. J. Soc. Magn. Reson. Med. Soc. Magn. Reson. Med. 2009;62:1619–1628.

[3] Letzen JE, Boissoneault J, Sevel LS, Robinson ME. Test-retest reliability of pain-related functional brain connectivity compared to pain self-report. Pain 2015.

[4] Nunnally JC. Introduction to Psychological Measurement. New York: McGraw Hill, 1970.

[5] Peacock S, Patel S. Cultural Influences on Pain. Rev. Pain 2008;1:6–9.

[6] Robinson ME, Staud R, Price DD. Pain Measurement and Brain Activity: Will Neuroimages Replace Pain Ratings? J. Pain 2013;14:323–327.

[7] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 1979;86:420–428.

[8] U. S. Food and Drug Administration/Center for Drug Evaluation and Research. Laboratory Manual. 2005.

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

[10] Woo C-W, Roy M, Buhle JT, Wager TD. Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. PLoS Biol. 2015;13:e1002036.

Figure 1

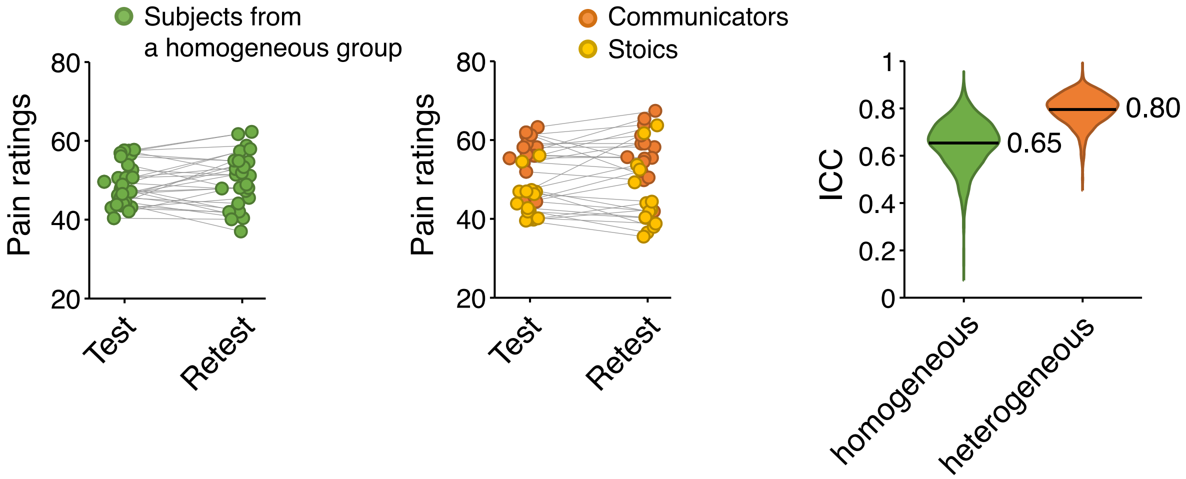
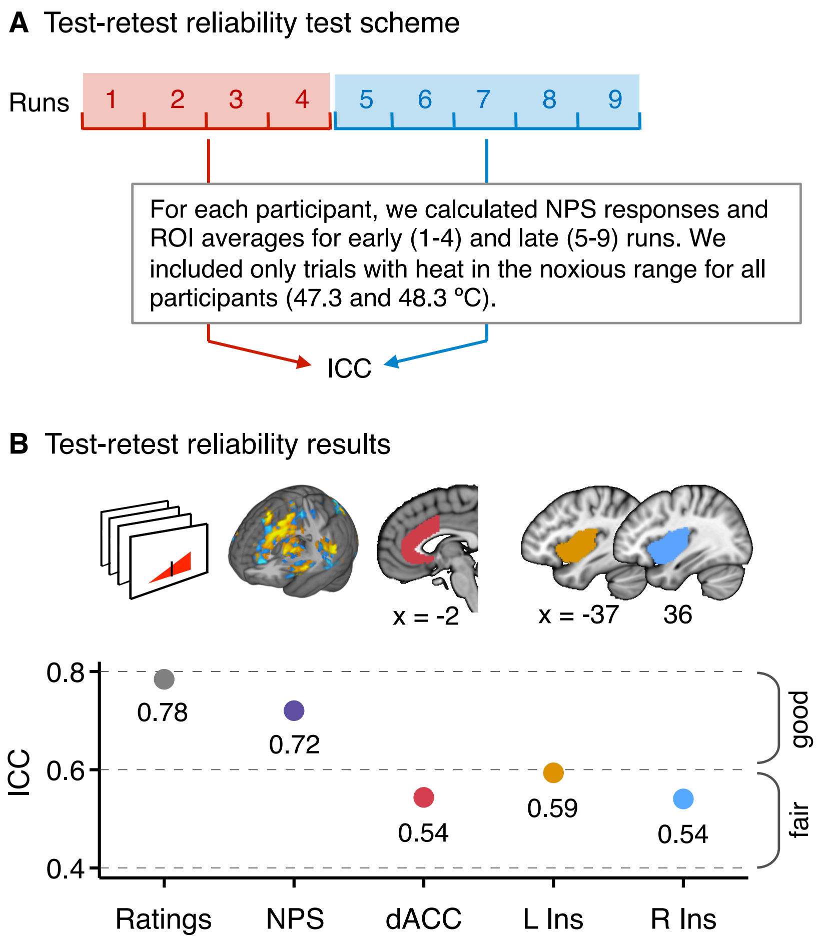

Subjects from
a homogeneous group
Pain ratings
80
60
40
20
Test
Retest
Communicators
Stoics
Pain ratings
80
60
40
20
Test
Retest
ICC
1
0.8
0.6
0.4
0.2
0
0.65
0.80
homogeneous
heterogeneous

Figure 2



A Test-retest reliability test scheme

For each participant, we calculated NPS responses and ROI averages for early (1-4) and late (5-9) runs. We included only trials with heat in the noxious range for all participants (47.3 and 48.3 °C).

B Test-retest reliability results