Reply

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
1–3
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797621996665
www.psychologicalscience.org/PS
SAGE

Need for Psychometric Theory in Neuroscience Research and Training: Reply to Kragel et al. (2021)

Maxwell L. Elliott[1], Annchen R. Knodt[1], Avshalom Caspi[1,2,3,4], Terrie E. Moffitt[1,2,3,4], and Ahmad R. Hariri[1]

[1]Department of Psychology & Neuroscience, Duke University; [2]Social, Genetic, & Developmental Psychiatry Research Centre, Institute of Psychiatry, Psychology, & Neuroscience, King's College London; [3]Department of Psychiatry & Behavioral Sciences, Duke University School of Medicine; and [4]Center for Genomic and Computational Biology, Duke University

Received 12/9/20; Revision accepted 1/14/21

We applaud the forward-looking nature of the Commentary provided by Kragel et al. (2021) on our article, "What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis" (Elliott et al., 2020). We fully agree with their emphasis on the importance of avoiding overgeneralization when considering measurement reliability in task-functional MRI (task-fMRI). Because no single reliability estimate can capture the multitude of possible task-fMRI measures, statements such as "every brain activity study you've ever read is wrong" (Cohen, 2020) are misleading and unnecessarily undermine our joint efforts to improve task-fMRI. In fact, that is why we addressed this very point in our article (see p. 801). Nevertheless, we take this opportunity to clarify three subtle but meaningful ways that our perspective diverges from that promoted by Kragel et al. in their Commentary.

First, as we embrace the future, we must also account for and build on the past while being realistic about the state of the present. Kragel et al. point out the exciting potential of "multivariate measures optimized using machine learning" that they claim are "commonly used for biomarker discovery" (p. XXX). While we agree that multivariate measures are becoming more widespread and should continue to be developed and explored (see p. 802 of our original article), such measures are still far from being universal in task-fMRI biomarker research. Because psychological science is a cumulative enterprise, criticism and honest assessment of the current state of the science are essential to the continued advancement of the field. In this vein, we surveyed the reliability of region-of-interest-based task-fMRI activation, which is one of the most commonly adopted measures reported in the literature over the past 2 decades. Our meta-analysis directly provided evidence for this continued use, as approximately half of the reliability estimates we found had been published in the previous 5 years. These measures are not relics of the past; they are in common use today and still frequently incorporated as primary measures in large-scale, state-of-the-art imaging efforts focused on biomarker development and individual-differences research. For example, the Human Connectome Project, UK Biobank, and the Adolescent Brain Cognitive Development study all have incorporated fMRI tasks designed to activate particular brain areas and circuits (Casey et al., 2018; Miller et al., 2016; Van Essen et al., 2013). These are large-scale, expensive projects creating MRI data sets for future neuroscience research. Thus, the poor reliability reported in our article is critical for not only past but also present biomarker research using traditional task-fMRI activation. We hope that by reevaluating standard practices in light of the reliability limitations detailed in our article, we can guard against repeating and perpetuating these limitations in such future research.

Second, we would like to highlight an important distinction between the main aim of our article and several of the examples offered by Kragel et al. in their Commentary. The central concern addressed in our article was whether commonly used measures of task-fMRI activation are reliable enough for individual-differences research and brain biomarkers. To answer this question,

Corresponding Author:
Ahmad R. Hariri, Department of Psychology & Neuroscience, Duke University
E-mail: ahmad.hariri@duke.edu

we assessed the test-retest reliability of task activation in the tradition of Cronbach's so-called correlational discipline of scientific psychology (Cronbach, 1957). However, Kragel et al. include examples of both between-subjects reliability per the correlational discipline and within-subjects replicability per the experimental discipline. For example, Figure S1c in Kragel et al.'s Supplemental Material demonstrates the replicability of machine-learning weights, across independent samples, that were used to classify faces and shapes across experimental conditions. While the ability of task-fMRI to decode experimental conditions may be of scientific interest, it is fundamentally a within-subjects experimental effect, falling within Cronbach's "experimental" discipline of scientific psychology. As we pointed out in our original article, "Within-subjects robustness is . . . often inappropriately invoked to suggest between-subjects reliability, despite the fact that reliable within-subjects experimental effects at a group level can arise from unreliable between-subjects measurements" (Elliott et al., 2020, p. 802). For example, contrasting faces with shapes consistently elicits amygdala activation within a group of individuals, despite poor test-retest reliability of the same amygdala activation between individuals (see http://haririlab.com/vid/ReliabilityTutorial.mp4). It is critical to preserve this often-confused distinction in order to ensure that reliability metrics are appropriately applied and interpreted within the research framework (Fröhner et al., 2019; Hedge et al., 2018).

Third, and related to the second point, we agree with Kragel et al. that different types of biomarkers require different demonstrations of reliability highlighted in their Figure 1a. However, we disagree with their description of COVID-19 tests (and diagnostic biomarkers more generally) as an example of a biomarker that does not need high test-retest reliability. In fact, a COVID-19 test desperately requires high test-retest reliability; however, it must be investigated over an appropriate timescale. Critically, COVID-19 tests must validly track changes in the underlying construct of interest (i.e., SARS-CoV-2 viral load). Therefore, when administered to an infected individual, a COVID-19 test should be capable of repeatedly returning positive results over minutes, hours, and even days until the moment that the disease status changes. More generally, although diagnostic biomarkers are naturally expected to change over time, they must be reliable within states (e.g., infected) so that deviations can be unambiguously attributed to a change in state (e.g., convalescence). Similarly, the time interval of reliability studies in task-fMRI should be calibrated to the putative timescale of stability within the underlying constructs of interest (e.g., hippocampal activity related to cognitive decline a year later). In hindsight, we should have more clearly stated in our original article that when we use the term "biomarker" we are referring to a specific class of between-subjects traitlike biomarkers useful for long-term prognostication.

In conclusion, we share the optimism of Kragel et al. about the potential for task-fMRI. In our own research, we have and will continue to enthusiastically work toward advancing reliable functional brain biomarkers. As other areas of fMRI (e.g., Elliott et al., 2019; Noble et al., 2017; Zuo et al., 2019) and biomedical research (e.g., Sugden et al., 2020) grapple with measurement challenges, we hope that further discussion of reliability in task-fMRI will similarly bear fruit in the form of reliable measures that help build a stronger, more cumulative science. Indeed, it is high time that neuroscience and psychometric theory come together in research, teaching, and training.

Transparency

Action Editor: John Jonides
Editor: Patricia J. Bauer
Author Contributions

All the authors contributed to the writing and revision of the manuscript. All the authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iDs

Maxwell L. Elliott https://orcid.org/0000-0003-1083-6277
Avshalom Caspi https://orcid.org/0000-0003-0082-4600
Ahmad R. Hariri https://orcid.org/0000-0003-3052-9880

References

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., Thomas, K. M., . . . Dale, A. M. (2018). The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, *32*, 43–54. https://doi.org/10.1016/j.dcn.2018.03.001

Cohen, A. (2020, June 25). *Duke University researchers say every brain activity study you've ever read is wrong*. https://www.fastcompany.com/90520750/duke-university-researchers-say-every-brain-activity-study-youve-ever-read-is-wrong

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684.

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. https://doi.org/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. *Psychological Science*, *31*(7), 792–806. https://doi.org/10.1177/0956797620916786

Fröhner, J. H., Teckentrup, V., Smolka, M. N., & Kroemer, N. B. (2019). Addressing the reliability fallacy in fMRI: Similar group effects may arise from unreliable individual effects. *NeuroImage*, *195*, 174–189. https://doi.org/10.1101/215053

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Kragel, P. A., Han, X., Kraynak, T. E., Gianaros, P. J., & Wager, T. D. (2021). Functional MRI can be highly reliable, but it depends on what you measure: A commentary on Elliott et al. (2020). *Psychological Science*, *32*, XXX–XXX. https://doi.org/10.1177/0956797621989730

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., . . . Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536. https://doi.org/10.1038/nn.4393

Noble, S., Spann, M. N., Tokoglu, F., Shen, X., Constable, R. T., & Scheinost, D. (2017). Influences on the test–retest reliability of functional connectivity MRI and its relationship with behavioral utility. *Cerebral Cortex*, *27*(11), 5415–5429. https://doi.org/10.1093/cercor/bhx230

Sugden, K., Hannon, E. J., Arseneault, L., Belsky, D. W., Corcoran, D. L., Fisher, H. L., Houts, R. M., Kandaswamy, R., Moffitt, T. E., Poulton, R., Prinz, J. A., Rasmussen, L. J. H., Williams, B. S., Wong, C. C. Y., Mill, J., & Caspi, A. (2020). Patterns of reliability: Assessing the reproducibility and integrity of DNA methylation measurement. *Patterns*, *1*(2), Article 100014. https://doi.org/10.1016/j.patter.2020.100014

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Zuo, X.-N., Xu, T., & Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nature Human Behaviour*, *3*, 768–771. https://doi.org/10.1038/s41562-019-0655-x