Commentary

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
2021, Vol. 32(4) 622–626
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797621989730
www.psychologicalscience.org/PS
SAGE

Functional MRI Can Be Highly Reliable, but It Depends on What You Measure: A Commentary on Elliott et al. (2020)

Philip A. Kragel[1], Xiaochun Han[2], Thomas E. Kraynak[3], Peter J. Gianaros[3], and Tor D. Wager[2]
[1]Department of Psychology, Emory University; [2]Department of Psychological and Brain Sciences, Dartmouth College; and [3]Department of Psychology, Center for the Neural Basis of Cognition, University of Pittsburgh

Received 7/7/20; Revision accepted 11/27/20

In a recent article, Elliott and colleagues (2020) evaluated the reliability of individual differences in task-based functional MRI (fMRI) activity and found reliability to be poor. They concluded that “commonly used task-fMRI measures generally do not have the test-retest reliability necessary for biomarker discovery or brain–behavior mapping” (p. 801). This is an important and timely effort, and we applaud it for spotlighting the need to evaluate the measurement properties of fMRI. Large samples combined with pattern-recognition techniques have made translational applications finally seem within reach. As the field gets serious about using the brain to predict behavior and health outcomes, reliability will become increasingly important.

However, along with their findings and constructive criticism comes the potential for overgeneralization. Though Elliott et al. focused on arguably the most limited fMRI measure for biomarker development—the average response within individual brain regions—the article has garnered media attention that mischaracterizes its conclusions. One headline reads, “every brain activity study you’ve ever read is wrong” (Cohen, 2020). The causes of anti-fMRI sentiments are not our concern here, but it is important to specify the boundary conditions of Elliott et al.’s critique. As they suggest and we show below, fMRI can exhibit high test-retest reliability when multivariate measures are used. These measures, however, were not evaluated by Elliot et al., despite being commonly used for biomarker discovery (Woo, Chang, et al., 2017). Thus, their conclusions do not apply to all “common task-fMRI measures” but to a particular subset that does not represent the state of the art. Moreover, there are multiple use cases for fMRI biomarkers (FDA-NIH Biomarker Working Group, 2016)—many of which do not require high test-retest reliability (cf. Elliott et al.; Fig. 1a).

Test-retest reliability estimates summarized by Elliott et al. reflect several limitations of the studies in their sample. These studies had (a) small sample sizes; (b) little data per participant (as little as 5 min); (c) single-task rather than composite-task measures, which can be more reliable (Gianaros et al., 2017; Kragel, Kano, et al., 2018); and (d) variable test-retest intervals, up to 140 days in the Human Connectome Project (HCP) data; in addition, they were limited to activity in individual brain regions.

Multivariate measures optimized using machine learning can have high test-retest reliability (Woo & Wager, 2016). Elliott et al. acknowledged this possibility, but did not provide quantitative examples. Examining some benchmarks from recent studies reveals that the situation is not nearly so dire for task-fMRI as Elliott et al. concluded. For example, Gianaros et al. (2020) identified patterns predictive of risk for cardiovascular disease using an emotional picture-viewing task and the HCP Emotion task. The same-day test-retest reliability of these measures was good to excellent (Spearman-Brown *r*s = .82 and .73, *N*s = 338 and 427, respectively; Fig. 1b). In contrast, test-retest reliabilities of individual regions (e.g., amygdala) were much lower (*r*s = .11–.27). In a second example, we assessed the same-day test-retest reliability of the neurologic pain signature—a neuromarker for evoked pain—in eight fMRI studies (*N* = 228; data from Geuter et al., 2020; Jepma et al., 2018). Reliability was good to excellent in all studies

Corresponding Authors:
Philip A. Kragel, Department of Psychology, Emory University
E-mail: pkragel@emory.edu

Tor D. Wager, Department of Psychological and Brain Sciences, Dartmouth College
E-mail: tor.d.wager@dartmouth.edu

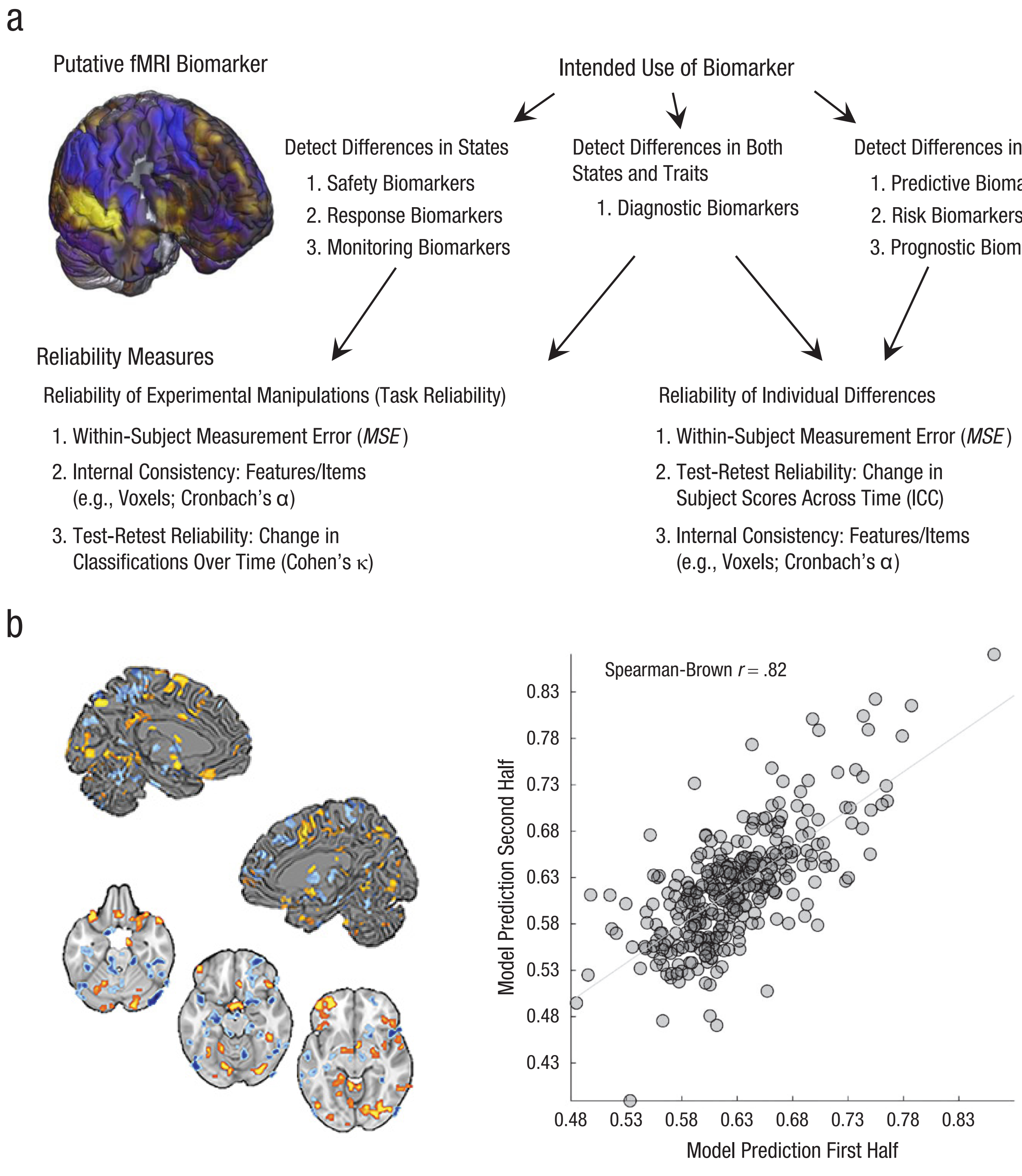
a
Putative fMRI Biomarker
Intended Use of Biomarker
Detect Differences in States
1. Safety Biomarkers
2. Response Biomarkers
3. Monitoring Biomarkers
Detect Differences in Both States and Traits
1. Diagnostic Biomarkers
Detect Differences in Traits
1. Predictive Biomarkers
2. Risk Biomarkers
3. Prognostic Biomarkers
Reliability Measures
Reliability of Experimental Manipulations (Task Reliability)
1. Within-Subject Measurement Error (MSE)
2. Internal Consistency: Features/Items (e.g., Voxels; Cronbach's α)
3. Test-Retest Reliability: Change in Classifications Over Time (Cohen's κ)
Reliability of Individual Differences
1. Within-Subject Measurement Error (MSE)
2. Test-Retest Reliability: Change in Subject Scores Across Time (ICC)
3. Internal Consistency: Features/Items (e.g., Voxels; Cronbach's α)
b
Spearman-Brown r = .82
0.83
0.78
0.73
0.68
0.63
0.58
0.53
0.48
0.43
Model Prediction Second Half
0.48
0.53
0.58
0.63
0.68
0.73
0.78
0.83
Model Prediction First Half

Fig. 1. Use cases for functional MRI (fMRI) biomarkers and an example of test-retest reliability. Among the seven major categories of biomarkers defined by the U.S. Food and Drug Administration (a), some (i.e., predictive, risk, and prognostic biomarkers) are designed to measure variation between individuals (e.g., to measure traitlike variables such as risk for depression, trait anxiety, or vulnerability to drug overdose). These biomarkers depend on measuring stable interindividual differences and thus require long-term test-retest reliability, which is typically estimated by calculating the intraclass correlation coefficient (ICC) for continuous variables or Cohen's κ for binary variables. Other biomarkers (i.e., safety, pharmacodynamic, monitoring, and response biomarkers) rely on the ability to measure variation within an individual across time, mental or physiological states, or treatment doses. Detecting within-person states relies less on stable individual differences than stable mappings between measure and state (in fMRI, between the brain and mental states and outcomes) with large and consistent effect sizes, referred to as *task reliability* (Hedge et al., 2018). This depends on low within-person measurement error (e.g., *MSE*) and can be measured with ICC or κ. For biomarkers related to dynamic states, other characteristics that increase test-retest reliability, including between-person heterogeneity and long-term stability across time, can be irrelevant or even undesirable. Reliability (b) is shown for a multivariate signature of risk for cardiovascular disease (figure adapted from Gianaros et al., 2020). The brain images depict significant pattern weights fitted on brain responses to affective images that positively (warm colors) and negatively (cool colors) contribute to the prediction of a marker of preclinical atherosclerosis. The scatterplot (with best-fitting regression line) depicts data used to estimate split-half reliability (*N* = 338).

(Fig. S1a in the Supplemental Material available online). Other multivariate measures show strong evidence for test-retest reliability across longer time intervals (Zuo & Xing, 2014). For example, Drysdale et al. identified four distinct fMRI biotypes for depression (total $N >$ 1,000).[1] Test-retest analyses across 4 weeks showed 90% agreement of biotype classifications (Fig. S1b in the Supplemental Material).

Functional MRI has promise for measuring individual differences, yet it may be best suited to develop biomarkers that detect dynamic states. Functional MRI patterns can reveal specific brain states—for example, whether a person is viewing a face (Haxby et al., 2001), replaying a memory (Momennejad et al., 2018), paying attention (Rosenberg et al., 2016), engaging in self-regulation (Cosme et al., 2020), or experiencing pain (Wager et al., 2013). This can be done reliably across long timescales. For example, Kamitani and Tong (2005) developed multivariate patterns that predict the orientation of a line viewed by participants. These patterns predicted line orientation 31 to 40 days after the initial prediction with similar performance (0.7 to 1 degree of error). As a second example, we reanalyzed HCP data presented by Elliott et al. and found that a multivariate model designed to identify the engagement of face (vs. shape) processing had excellent reliability across 4 weeks (Fig. S1c in the Supplemental Material). These examples and many others (Zuo & Xing, 2014, Finn et al., 2015) show that fMRI can yield measures that reliably detect the emergence of brain states across time.

A common belief is that all biomarkers measure traits and thus require high test-retest reliability, but we argue that this is a misconception. The U.S. Food and Drug Administration identifies seven major categories of biomarkers (Fig. 1a). Some, such as *prognostic* biomarkers for future disease and *predictive* biomarkers for treatment response, rely on individual differences in stable traits and require long-term test-retest reliability. Others measure states and require low within-person measurement error but not necessarily high test-retest reliability (Bland & Altman, 1996). For example, *diagnostic* biomarkers for disease states—such as COVID-19—need not be stable across weeks to months as the disease state changes. The same is true for *safety*, *monitoring*, and *pharmacodynamic* biomarkers that track changes in pathophysiological states. Measures of fMRI show promise in detecting and monitoring disorder-related brain processes (Duff et al., 2020; Rosenberg et al., 2016; Woo, Schmidt, et al., 2017). Although a full discussion of use cases is beyond the scope of this Commentary (but see Davis et al., 2020), fMRI measures can be sufficiently sensitive, specific, and reliable for many uses.

Ultimately, reliability is not a fixed property of an assay, let alone a whole measurement technology. It depends on the tasks, samples, and measures extracted from them (Streiner, 2003). Other criticisms of fMRI (Eklund et al., 2016; Vul et al., 2009) have been used to issue blanket condemnations. We caution against such overgeneralization and propose that the findings of Elliott et al. be considered a lower bound on the reliability of fMRI. The upper bound is high and remains to be fully explored. We agree with the summary recommendations for future fMRI research made by Elliott et al. and are optimistic that new methods designed to optimize reliability (Dubois & Adolphs, 2016) while keeping construct validity in mind (Kragel, Koban, et al., 2018) will continue to fuel fMRI research on biomarker development.

Transparency

Action Editor: John Jonides
Editor: Patricia J. Bauer
Author Contributions
P. A. Kragel and T. D. Wager conceived the project and data-analysis plan. P. A. Kragel, X. Han, and T. E. Kraynak performed the analyses. P. A. Kragel and T. D. Wager wrote the manuscript. T. D. Wager and P. J. Gianaros designed, implemented, and oversaw data collection and generation of the research protocol. All the authors contributed to the revision of the manuscript and 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.
Funding
This project was supported by Grants R01 MH076136, DA046064, MH116026, EB026549, and P01 HL040962.
Open Practices
Data and materials for this study have not been made publicly available, and the design and analysis plans were not preregistered.

ORCID iD

Philip A. Kragel https://orcid.org/0000-0001-9463-6381

Acknowledgments

We thank Bogdan Petre for assistance with data analysis.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797621989730

Note

[1] A recent study failed to reproduce the biotype-development process in a different sample (Dinga et al., 2019) but did not test Drysdale et al.'s biotypes.

References

Bland, J. M., & Altman, D. G. (1996). Measurement error. *The BMJ*, *312*(7047), Article 1654. https://doi.org/10.1136/bmj.312.7047.1654

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

Cosme, D., Zeithamova, D., Stice, E., & Berkman, E. T. (2020). Multivariate neural signatures for health neuroscience: Assessing spontaneous regulation during food choice. *Social Cognitive and Affective Neuroscience*, *15*(10), 1120–1134. https://doi.org/10.1093/scan/nsaa002

Davis, K. D., Aghaeepour, N., Ahn, A. H., Angst, M. S., Borsook, D., Brenton, A., Burczynski, M. E., Crean, C., Edwards, R., Gaudilliere, B., Hergenroeder, G. W., Iadarola, M. J., Iyengar, S., Jiang, Y., Kong, J.-T., Mackey, S., Saab, C. Y., Sang, C. N., Scholz, J., . . . Pelleymounter, M. A. (2020). Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: Challenges and opportunities. *Nature Reviews Neurology*, *16*, 381–400.

Dinga, R., Schmaal, L., Penninx, B. W. J. H., van Tol, M. J., Veltman, D. J., van Velzen, L., Mennes, M., van der Wee, N. J. A., & Marquand, A. F. (2019). Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017). *NeuroImage*, *22*, Article 101796. https://doi.org/10.1016/j.nicl.2019.101796

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R. N., . . . Liston, C. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, *23*(1), 28–38.

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends in Cognitive Sciences*, *20*(6), 425–443.

Duff, E. P., Moultrie, F., van der Vaart, M., Goksan, S., Abos, A., Fitzgibbon, S. P., Baxter, L., Wager, T. D., & Slater, R. (2020). *Inferring the infant pain experience: A translational fMRI-based signature study. bioRxiv*. https://doi.org/10.1101/2020.04.01.998864

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences, USA*, *113*(28), 7900–7905.

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

FDA-NIH Biomarker Working Group. (2016). *BEST (biomarkers, endpoints, and other tools) resource*. Food and Drug Administration.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*(11), 1664–1671.

Geuter, S., Losin, E. A. R., Mathieu Roy, A., Lauren, Y., Schmidt, L., Krishnan, A., Koban, L., Wager, T. D., & Lindquist, M. A. (2020). Multiple brain networks mediating stimulus–pain relationships in humans. *Cerebral Cortex*, *30*(7), 4204–4219.

Gianaros, P. J., Kraynak, T. E., Kuan, D. C.-H., Gross, J. J., McRae, K., Hariri, A. R., Manuck, S. B., Rasero, J., & Verstynen, T. D. (2020). Affective brain patterns as multivariate neural correlates of cardiovascular disease risk. *Social Cognitive and Affective Neuroscience*, *15*(10), 1034–1045. https://doi.org/10.1093/scan/nsaa050

Gianaros, P. J., Sheu, L. K., Uyar, F., Koushik, J., Jennings, J. R., Wager, T. D., & Verstynen, T. D. (2017). A brain phenotype for stressor-evoked blood pressure reactivity. *Journal of the American Heart Association*, *6*(9), Article e006053. https://doi.org/10.1161/JAHA.117.006053

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

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.

Jepma, M., Koban, L., van Doorn, J., Jones, M., & Wager, T. D. (2018). Behavioural and neural evidence for self-reinforcing expectancy effects on pain. *Nature Human Behaviour*, *2*(11), 838–855.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685.

Kragel, P. A., Kano, M., Van Oudenhove, L., Ly, H. G., Dupont, P., Rubio, A., Delon-Martin, C., Bonaz, B. L., Manuck, S. B., Gianaros, P. J., Ceko, M., Reynolds Losin, E. A., Woo, C.-W., Nichols, T. E., & Wager, T. D. (2018). Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. *Nature Neuroscience*, *21*(2), 283–289.

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

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, *7*, Article e32548. https://doi.org/10.7554/elife.32548

Rosenberg, M. D., Finn, E. S., Scheinost, D., Papademetris, X., Shen, X., Constable, R. T., & Chun, M. M. (2016). A neuromarker of sustained attention from whole-brain functional connectivity. *Nature Neuroscience*, *19*(1), 165–171.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. https://doi.org/10.1111/j.1745-6924.2009.01125.x

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. *The New England Journal of Medicine*, *368*(15), 1388–1397.

Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience*, *20*(3), 365–377.

Woo, C.-W., Schmidt, L., Krishnan, A., Jepma, M., Roy, M., Lindquist, M. A., Atlas, L. Y., & Wager, T. D. (2017). Quantifying cerebral contributions to pain beyond nociception. *Nature Communications*, *8*, Article 14211. https://doi.org/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*(3), 511–513. https://doi.org/10.1097/j.pain.0000000000000442

Zuo, X. N., & Xing, X. X. (2014). Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: A systems neuroscience perspective. *Neuroscience & Biobehavioral Reviews*, *45*, 100–118. https://doi.org/10.1016/j.neubiorev.2014.05.009