Cogn Affect Behav Neurosci
DOI 10.3758/s13415-013-0201-7

EDITORIAL

Introduction to the special issue on reliability and replication in cognitive and affective neuroscience research

Deanna M. Barch · Tal Yarkoni

Researchers in many areas of psychology and neuroscience have grown concerned with what has been referred to as a "crisis" of replication and reliability in the field. These concerns have cut across a broad range of disciplines and have been raised in both the biomedical (Ioannidis, 2005, 2011) and psychological (Pashler & Harris, 2012; Simmons, Nelson, & Simonsohn, 2011) sciences. A number of reasons have been put forth for these concerns about replication, including conflicts of interest (Bakker & Wicherts, 2011; Ioannidis, 2011), misaligned incentives, questionable research practices (John, Loewenstein, & Prelec, 2012) that result in what has been referred to as "p-hacking" (Simmons et al., 2011), and ubiquitous low power (Button et al., 2013). Such problems lead to the publication of inflated effect sizes (Masicampo & Lalande, 2012; Vul & Pashler, 2012) and produce a higher incidence of false positives.

One could read this emerging literature and walk away disheartened by the state of many scientific fields—and perhaps of science in general. Alternatively, one could take this as an opportunity to step back and develop new procedures and methods for tackling at least some of the problems contributing to the crisis of replication, whether real or perceived. That is what is attempted in this special issue of *Cognitive, Affective, & Behavioral Neuroscience*, with a specific emphasis on studies that use functional neuroimaging to understand the neural mechanisms that support a range of cognitive and affective processes. The articles in this special issue fall into three general categories: (1) research into the importance and influence of methods choices and reporting; (2) assessments of reliability and novel approaches to statistical analysis; and (3) studies of power, sample size, and the importance of both false positives and false negatives.

In terms of methodological concerns, it is important to note that studies using functional neuroimaging methods to study cognitive and affective processes need to be concerned with all of the same issues that apply to any behavioral study of a psychological process. As Plant et al. describe in their article in this issue, concerns about the timing of stimulus presentation and response collection in studies of psychological processes may contribute to replication difficulties, and these authors suggest some ways to assess this possibility and potentially correct it. In addition, studies in cognitive and affective neuroscience are subject to the same concerns about transparency in the reporting of methods as are behavioral studies, including all of the problematic behaviors that Simmons and others have suggested lead to spurious rejection of the null hypothesis (Simmons et al., 2011). These concerns led Simmons and colleagues to propose that all authors be required to include the following 21-word phrase in their Method sections: "We report how we determined our sample size, all data exclusions (if any), all manipulations and all measures in the study" (Simmons, Nelson, & Simonsohn, 2012). Although the inclusion of such a statement has yet to be widely adopted by journals, it highlights the critical issue of transparency in method reporting. One of the major concerns in the field of functional neuroimaging—whether reporting functional magnetic resonance imaging (fMRI), event-related potentials (ERP), transcranial magnetic stimulation (TMS), or other techniques—is the plethora of analysis choices that a researcher can make, and the influence that these different choices clearly have on the results. Poldrack outlined this concern in his earlier work and suggested guidelines for reporting method choices that could influence outcomes (Poldrack et al., 2008). In another prior work, Carp (2012) reviewed and summarized the failure of many researchers to follow these guidelines or to report on key methodological

D. M. Barch (✉)
Departments of Psychology, Psychiatry, and Radiology, Washington University, St. Louis, MO, USA
e-mail: dbarch@wustl.edu

T. Yarkoni
Department of Psychology, University of Colorado, Boulder, CO, USA

Published online: 07 August 2013

details that could influence replication. In the present issue, Carp reiterates the importance of detailed method reporting, and extends the prior work by providing suggestions for new ways to share processing streams and analysis codes that may help enhance replicability and identify differences in method choices that could influence outcomes. The present article by Carp focuses on these method reporting issues in relationship to fMRI research, but the same concerns apply to other types of cognitive neuroscience methods. As such, the Society for Psychophysiological Research is also currently working on guidelines for improved method report for ERP and electroencephalography (EEG) research.

Another important area for enhancing replication is in the domain of improved statistical analyses that might provide clearer information about replicability and enhanced robustness of results. For example, in the present issue, Coutanche reviews the use of multivariate pattern analysis (MVPA) as an alternative approach to standard univariate analyses for extracting information from fMRI. The articles by Lindquist, Poppe et al., Bennett and Miller, Tuner and Miller, and Weber et al. focus on ways to assess and enhance the reliability of varying approaches to statistical analysis of fMRI data. Lindquist proposes a novel image-based intraclass correlation approach (referred to as *I2C2*) to examining replication, along with a bootstrapping method for assessing the variability in the I2C2 and a Monte Carlo permutation approach to assess the degree to which the observed reliability is significantly different from zero. Using an image-based approach to assess reliability has some disadvantages (as opposed to a region- or component-based approach), as reliability can vary considerably across brain regions, depending on their relative involvement in a task as well as on a number of other factors (see Bennett & Miller's article in this issue for some examples). However, this image-based approach does offer a way to provide a global assessment of reliability, and it would be particularly appropriate for domains in which the patterns of activity or connectivity across all or much of the brain are the key dependent variables of interest (e.g., MVPA, patterns of connectivity, etc.).

In other work focused on measurement and improvement of reliability, Poppe et al. investigate the reliability of independent component analysis (ICA) during task and rest, and assess the impacts of several methodological choices (i.e., number of subject orders, dimensionality, and threshold) on reliability. Replication within additional cognitive or affective domains would enhance the generalizability of these results, though Poppe et al. did use two independent data sets and both task and resting-state components. Bennett and Miller also examine a range of key factors that could influence the reliability of analyses, by using a general linear model (GLM) approach. These factors included cognitive domain, type of task design, type of contrast, and the nature of the statistical threshold. Although it remains an open question to what extent the results that Bennett and Miller report—for example, that working memory had overall higher reliability than episodic memory—will generalize beyond their particular study, the important overarching message that they convey is that researchers should be examining the reliability of their particular task and analysis choices on a much more regular basis. In a related article, Turner and Miller focus on how reliability is influenced by the number of events available for a particular cell in an experimental design, and highlight the challenges of fMRI studies using events that are defined a posteriori by the behavior of the participant and that are not fully under experimental control. Weber et al. focus on the reproducibility of measures from graph-theory-based measures of networks and examine how factors such as the type of data (task vs. resting), the type of pulse sequence (BOLD vs. other), and the frequency range examined influence reproducibility. Together, the data from Poppe et al., Bennett and Miller, Turner and Miller, and Weber et al. provide concrete guidance for a range of study design choices that are most likely to enhance the reliability of several different approaches to examining brain activity and brain connectivity using general linear model, ICA, and graph-theory-based approaches to analysis. Furthermore, these articles point to the importance of a regular assessment of the reliably specific to the cognitive/affective domain, task design, and analysis approach used in a particular study.

A further interesting angle on improving statistical analyses is highlighted in the article by Eklund et al. on the ways in which the use of graphics processing units (GPUs) could enhance cognitive neuroscience methods, by virtue of allowing the use of computationally complex but potentially advantageous statistical methods. Some of the examples that they provide are improved spatial normalization methods and the use of Bayesian models, but the use of GPUs in general opens up a relatively inexpensive approach to enhanced computational power, if coding approaches can be developed that utilize GPUs.

A last, but critically important, domain addressed by articles in this special issue concerns sample size, power, and the balancing of false positives and false negatives. As has recently been quantified by Button et al. (2013), a major problem for the field of neuroimaging is the near-ubiquitous use of small sample sizes—and the concomitant low power—that plagues many studies, both published and unpublished. Unfortunately, a mythos has arisen around sample size for fMRI, so that many investigators believe that a fixed—and typically low—number of subjects provides sufficient power to detect effects, irrespective of the nature of the study and the phenomena under investigation. This belief is demonstrably incorrect and has arisen from several sources. One source is early studies on power that suggested that sample sizes in the range of 12–20 subjects were sufficient to achieve adequate power if a very liberal statistical threshold was used (Desmond &

Glover, 2002). More recently, Friston argued that sample sizes in the range of 16 to 32 should be enough for fMRI studies (Friston, 2012). Some factors that such arguments fail to consider are that more sophisticated and well-controlled contrasts typically lead to small effect sizes; that many studies use between-group designs that have considerably lower power; and that careful control for multiple comparisons will increase the necessary sample size.

More generally, the use of small samples coupled with a very large number of comparisons (in whole-brain analyses) is likely to dramatically bias the results that one observes. Because of low power, effects are likely to seem more spatially selective than they truly are, and because of sampling error, those few effects lucky enough to survive correction are likely to have grossly inflated effect size estimates (Yarkoni, 2009). In the present issue, Durnez et al. address the former of these issues, focusing on the likelihood of false negatives, or failing to detect true effects because of insufficient power. They put forth a novel take on this issue, developing and evaluated a statistical threshold approach that provides information about both false positives and false negatives. Mar et al. have adopted a complementary approach, arguing for the importance of data aggregation across investigators and studies as a way to enhance sample size and power, with a particular focus on personality neuroscience. Although there are likely to be significant ethical and logistic limitations on such an approach, the general idea of building bridges across data sets is a critical one that needs careful consideration and implementation in the field. Of course, efforts to improve power—whether through the introduction of novel statistical procedures or aggregation across data sets—should not be taken as a license to use small sample sizes. Durnez et al.'s prescriptions are most useful in situations in which sample sizes are limited by factors beyond the control of the experimenter (presurgical MRI is the example used), and Mar et al.'s proposal is a means of compensating for extant limitations in the literature, not a justification for continued conventional practices of insufficient sample sizes.

Naturally, the articles in this special issue do not address all of the issues associated with the replication concerns noted in many scientific domains, nor are they a panacea for all challenges that face our field. However, they do try to tackle a number of the most critical and pressing issues, and many of the authors offer concrete and practical approaches that researchers can begin to integrate into their own work to enhance methodological clarify, reliability, and power, as a way to increase the robustness and interpretability of results reported in the field. We hope that these articles will spur researchers to think deeply about these issues for their own work, will provide them with some concrete tools to add to their arsenal, and will help drive the field to a next stage of cognitive and affective neuroscience research.

References

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43,* 666–678. doi:10.3758/s13428-011-0089-5

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. doi:10.1038/nrn3475

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage, 63,* 289–300. doi:10.1016/j.neuroimage.2012.07.004

Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods, 118,* 115–128. doi:10.1016/S0165-0270(02)00121-8

Friston, K. (2012). Ten ironic rules for non-statistical reviewers. *NeuroImage, 61,* 1300–1310. doi:10.1016/j.neuroimage.2012.04.018

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2011, May 31). An epidemic of false claims: Competition and conflicts of interest distort too many medical findings. *Scientific American, 304,* 16.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524–532. doi:10.1177/0956797611430953

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology, 65,* 2271–2279.

Pashler, H. E., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536.

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *NeuroImage, 40,* 409–414. doi:10.1016/j.neuroimage.2007.11.048

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Dialogue, 26,* 1. doi:10.2139/ssrn.2160588

Vul, E., & Pashler, H. (2012). Voodoo and circularity errors. *NeuroImage, 62,* 945–948. doi:10.1016/j.neuroimage.2012.01.027

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al. (2009). *Perspectives on Psychological Science, 4,* 294–298. doi:10.1111/j.1745-6924.2009.01127.x