doi:10.1093/scan/nsp052

Tools of the Trade

Type I and Type II error concerns in fMRI research: re-balancing the scale

Matthew D. Lieberman,[1] **and William A. Cunningham**[2]

[1]Departments of Psychology, Psychiatry, & Biobehavioral Sciences, University of California, Los Angeles, and [2]Department of Psychology, Ohio State University, USA

Statistical thresholding (i.e. *P*-values) in fMRI research has become increasingly conservative over the past decade in an attempt to diminish Type I errors (i.e. false alarms) to a level traditionally allowed in behavioral science research. In this article, we examine the unintended negative consequences of this single-minded devotion to Type I errors: increased Type II errors (i.e. missing true effects), a bias toward studying large rather than small effects, a bias toward observing sensory and motor processes rather than complex cognitive and affective processes and deficient meta-analyses. Power analyses indicate that the reductions in acceptable *P*-values over time are producing dramatic increases in the Type II error rate. Moreover, the push for a mapwide false discovery rate (FDR) of 0.05 is based on the assumption that this is the FDR in most behavioral research; however, this is an inaccurate assessment of the conventions in actual behavioral research. We report simulations demonstrating that combined intensity and cluster size thresholds such as $P < 0.005$ with a 10 voxel extent produce a desirable balance between Types I and II error rates. This joint threshold produces high but acceptable Type II error rates and produces a FDR that is comparable to the effective FDR in typical behavioral science articles (while a 20 voxel extent threshold produces an actual FDR of 0.05 with relatively common imaging parameters). We recommend a greater focus on replication and meta-analysis rather than emphasizing single studies as the unit of analysis for establishing scientific truth. From this perspective, Type I errors are self-erasing because they will not replicate, thus allowing for more lenient thresholding to avoid Type II errors.

Keywords: social cognitive neuroscience; MR statistics; type II error

Almost every fMRI analysis involves thousands of simultaneous significance tests on discrete voxels in collected brain volumes. As a result, setting one's *P*-value threshold to 0.05, as is typically done in the behavioral sciences, is sure to produce hundreds or thousands of false positives in every analysis. To guard against such errors, MR statisticians have worked to develop methods that are increasingly effective at guarding against Type I errors (i.e. false positives resulting from noise when there is no true effect). Clusters of activity that survive these methods are thought to reflect true effects and thus are expected to replicate in future studies. The benefits of preventing Type I errors with these methods is undeniable, but precious little ink has been spilled or mental effort spent on considering the costs of a single-minded focus on Type I errors or whether our procedures, rather than emulating the actual statistical practices of the behavioral sciences, have considerably overshot the mark.

In this article, we focus on the primary cost of decreasing the likelihood of Type I errors, namely the inherent increase in another inferential error—the Type II error. The Type II error occurs when we fail to recognize a true effect. We argue here that in neuroimaging, relative to behavioral research, the Type II error may be actually the more pernicious of the two errors, because in neuroimaging, results that may constitute Type II errors are not reported at all because they do not meet the threshold for significance and cannot be considered in aggregate analyses such as meta-analyses. In addition, the measures taken to avoid Type I errors in neuroimaging are disproportionately likely to increase the number of Type II errors for tasks that allow for multiple cognitive solutions to a problem (e.g. decision-making tasks) and/or variability across trials in degree or timing of cognitive processing (e.g. emotional regulation in response to a particular stimulus).

A second focus of this article is the putative goal of MR statisticians of emulating the *P*-value conventions established and used by behavioral researchers. Although the official statistical norm of the behavioral sciences is to use $P < 0.05$ corrected for the number of tests conducted, this is not the convention used in almost any behavioral science publication. Most articles report many tests, each held to a standard

Received 12 January 2009; Accepted 12 September 2009

We would like to thank Elliot Berkman, Jeanette Mumford, Kevin Ochsner and Kris Preacher for discussion of the ideas in this article.

Correspondence should be addressed to Matthew Lieberman, Department of Psychology, Franz Hall, University of California, Los Angeles, CA 90095-1563, USA. E-mail: lieber@ucla.edu

of $P < 0.05$ without any correction for multiple tests. We agree that fMRI research should emulate the standards of behavioral research, but the standards of research as it is actually conducted. For a young science to impose stricter self-standards than its parent discipline, when doing so would increase Type II errors and lead us to miss important neural landmarks for future investigations, is not wise.

Cognitive neuroscience in general, and social and affective neuroscience in particular, is still in an exploration phase and thus it is more important not to dismiss possible true effects than it is to avoid reporting false alarms. We argue for a view of cognitive neuroscience that focuses more on the aggregation of data across multiple studies and in meta-analyses. From this perspective, false alarms are important but 'far' from catastrophic as they will not replicate and thus are self-erasing, but unreported Type II errors cannot be included in the multi-study aggregation of findings and can never be the basis for attempted replications or extensions.

AVOIDING TYPE I ERRORS

In the early days of fMRI and for quite some time, the gold standard for analysis was to use a *P*-value of 0.001 at each voxel. The thought was that in the absence of knowing what the correct *P*-value should be, using one that is 50 times more stringent than the one used by behavioral scientists is reasonable. And why not, given that $P < 0.05$ was a relatively arbitrary standard set by Fisher (1926) who wrote 'it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials"' (p. 504). Although countless classic fMRI studies have been conducted with the $P < 0.001$ threshold since the early 1990s, studies that have been replicated time and time again (e.g. Wagner *et al.*, 1998), increasingly this threshold is being discarded by groups seriously concerned with Type I errors. The reason seems simple: with the number of simultaneous tests conducted in an analysis, using $P < 0.001$ is still likely to produce up to 100 voxels (assuming an extreme of 100 000 voxels tested at once) that appear significant while in fact being false alarms.

A variety of statistical procedures have been introduced to better cope with the multiple comparisons issue in fMRI. One solution has been to jointly use intensity and spatial extent thresholds in the same analysis ('intensity/cluster thresholding'; Forman *et al.*, 1995). In such analyses, a researcher specifies a *P*-value that a voxel must surpass to be considered but also specifies the number of contiguous voxels (i.e. spatial extent) that must all surpass this *P*-value to be considered a significantly active cluster of voxels. The reasoning is that voxels representing false alarms due to noise will be randomly distributed throughout the brain and thus are much less likely to occur in contiguous groups of voxels than in single voxels. Using a slightly more liberal *P*-value ($P < 0.005$) along with a 10 voxel extent threshold is more likely to produce replicable data than a *P*-value of 0.001 alone, if other scanning parameters are held constant.

Recent approaches that are more conservative with respect to Type I errors are family-wise error (FWE) correction procedures and false discovery rate (FDR) techniques (Genovese *et al*, 2002). FWE approaches based on Monte Carlo simulation and/or Gaussian field theory using various cluster size corrections are conceptually similar to Bonferroni. In contrast, formal FDR analyses take the distribution of *P*-values from an analysis into account and end up being less conservative than FWE (yet still more likely to avoid Type I errors than intensity/cluster thresholding). While FWE analyses use a correction meant to result in almost no Type I errors, FDR techniques limit the rate of Type I errors and this difference is clearly a nod to Type II error concerns.[1]

THE COSTS OF SINGLE-MINDED FOCUS ON TYPE I ERRORS

The benefit of a nearly exclusive focus on Type I errors is clear. Using the most conservative of these methods (FWE), few Type I errors will occur in a given fMRI analysis and as a result, most observed effects will reflect true population level differences that will replicate in most sufficiently powered fMRI samples.

The costs of this focus are less clear and certainly less often considered. Here, we will consider four negative consequences: (i) increased Type II errors, (ii) a bias toward publishing large and obvious effects, (iii) a bias against observing effects associated with complex cognitive and affective processes and (iv) deficient meta-analyses. Although some of these consequences affect all fMRI studies equally, some of these disproportionately impact social and affective neuroscience investigations.

Increased Type II errors

Types I and II error rates are a zero-sum game for any given sample size. Any method that protects more against one type of error is guaranteed to increase the rate of the other kind of error. Consider a test that returns a *P*-value of 0.02. If this was one of 10 tests and we applied a Bonferroni correction, we would conclude that the null hypothesis could not be rejected. If all 10 tests had *P*-values of 0.02, we would end up rejecting all of the tests even though it is likely that several of the tests represent real effects. Thus, when we set increasingly

[1] It is interesting to note that many researchers who used $P < 0.005$ (10 voxels) or $P < 0.001$ (no extent threshold) for years have moved on to using FDR techniques. It is not uncommon for these researchers to suggest that new data should not be published unless it is using FDR correction to a 0.05 level. One would assume that such researchers would either want to retract their own previous work or at least publish corrigenda indicating that their previous results should not be taken seriously. One rejoinder would be the success with which the original work replicated and indeed, this demonstrates two of our main points. If their original work had not been published due to more conservative thresholding procedures, no one might have thought to replicate those effects. Moreover, the fact that these less conservative thresholding procedures produced now classic replicable effects is just one more indication that these thresholds can operate effectively. We hope it is obvious that we do not think such papers are in need of retraction or corrigenda.

conservative *P*-values, we are reducing the number of false alarms but we are also increasing the number of real effects that are effectively being treated as spurious.

The best estimate of Type II error rates comes from power analyses. There is no way to estimate, *a priori*, how many Type II errors occurred in an existing fMRI data set. However, power analyses do estimate the likelihood of a Type II error in future samples given a true effect of a certain size. Assume there is a real effect equivalent to $R^2 = 0.25$, considered a large effect in the behavioral sciences, and a sample size of 20. If we set our *P*-value threshold to 0.005, our likelihood of detecting this effect is 30.7%. In other words, given a real effect, our Type II error rate is 69.3%. Remember that this is using the liberal threshold that is advised against by those focused on Type I errors. Applying a *P*-value threshold of 0.001 and assuming the same true effect, our Type II error rate is 86.2%, but this threshold is still considered too liberal. Moving to a *P*-value threshold of 0.0001 and assuming the same true effect, our Type II error is 96.5%. Thus, steps taken to lower the FDR can have devastating effects on our likelihood of detecting sizable true effects.

Bias toward large obvious effects

Everyone knows the old story of the man who is looking for his keys at night under the streetlight. When asked if he dropped them under the streetlight he says 'No, but this is where the light is good'. As the balance of Types I and II conventions shift, it can have a similar effect on 'where the light is good' in terms of what kinds of fMRI investigations are likely to bear fruit. In the absence of constraints, we would all run samples with $n = 500$ and the concerns about both kinds of errors would largely vanish, but given the expense of neuroimaging, only the most successful labs could ever dream of running such large samples. Thus, with the constraint that most early to mid-career investigators will be able to scan samples of 16–20, power analyses dictate that as *P*-value thresholds are made more conservative, the only effects that can be examined with a decent likelihood of successful detection are very large effects. Securing funding to examine subtle effects in large samples with sufficient statistical power presents something of a Catch-22; in order to secure funding, one needs to present promising pilot data, but the expense of obtaining good data for subtle effects is precisely why the funding is needed. Paradoxically, only those studying large effects can get the pilot data in a small sample to secure the funding to collect large samples, which are not actually needed to study large effects. To be very clear about this, with $n = 20$ only effects with a Cohen's d of 1.85 or larger will identified as significant using $P < 0.001$ (and this is now considered an overly liberal significance threshold). Behavioral scientists all know that very important and interesting effects occur well below that effect size threshold.

Bias against complex cognitive and affective effects

For multiple reasons, fMRI studies examining sensory and motor phenomena are likely to have larger effect sizes than social and affective phenomena and thus are more resistant to measures designed to avoid Type I errors. One reason for this is that the regions of the brain involved in sensory and motor phenomena are largely immune to susceptibility artifacts. In contrast, multiple regions associated with social and affective phenomena such as ventromedial PFC, amygdala, and the temporal poles are more difficult to image because of susceptibility artifacts. Thus, just in terms of ability to extract signal with fMRI, social and affective neurosciences are at a disadvantage. Perhaps more significantly, there is a tighter mapping between experimental inputs and outcomes in sensory and motor domains than in the social and affective domains. A checkerboard pattern presented multiple times to different participants will have a very similar set of effects from trial to trial and from person to person, greatly enhancing the signal-to-noise ratio. Similarly, instructions to tap or not tap one's fingers will produce very reliable behavioral outputs. Critically, in both of these cases, the experimenter has precise timing information in terms of visual inputs or motor outputs that allows for precise modeling of the neural events. Given how central the notion of subjective construal is to social and affective phenomena, there is greater inherent trial-to-trial and person-to-person variability in these domains (Griffin and Ross, 1991). Forming a judgment of another person or reappraising one's emotions can invoke a variety of different processes and recruit different representational content depending on a perceiver's current and chronically accessible constructs, associations, and expectations (see Ochsner, 2007). Moreover, it is essentially impossible to gain the same precision timing over when the reappraisal process is occurring. This makes it a fundamentally noisier process to model. Over time, one would naturally want to be able to assess the different sources of variability (at least those that can be assessed) and examine the neural correlates of different kinds of responses, but if correction procedures are sufficiently severe, it might be difficult to ever find initial neural landmarks worth investigating further.

Deficient meta-analyses

One can debate the merits of adjusting our Type I focus for any of the previous costs described. One can reasonably argue that these are simply crosses that social and affective neuroscience must bear if they want to be at the table with the 'big boys'. But the final cost of a single-minded focus on Type I errors concerns anyone who uses fMRI. In behavioral science papers, *t*-values are typically reported whether the associated *P*-value is significant or not, whereas in typical fMRI papers, *t*-values (or their equivalent) are only reported if the *P*-value meets the criteria set for significance.

Meta-analyses serve a number of different purposes. First, any science is a cumulative endeavor and the results from any particular study need to be seen in light of similar studies. For example, if a result is seen only once, this particular finding (even if FDR corrected) will be disregarded to the extent that it does not replicate. Thus, Type I errors are defended against through the accumulation of data across studies and labs. Second, perhaps more relevant to the issue of Type II errors, meta-analyses are commonly used to detect effects that may not meet conventional significance levels in individual studies but are real and emerge when the studies are considered together in the aggregate. In other words, meta-analyses can compensate for the low power to detect subtle but real effects in small sample studies by leveraging their combined sample sizes. If the true effect size for a behavioral effect is $r = 0.10$, sample sizes of 30–40 will only rarely detect this as significant at $P < 0.05$. Nevertheless, as long as the *t*- or *P*-value is reported in each study, despite its within study non-significance, when 20 or more similar studies are combined meta-analytically, the effect may well emerge as significant.

Because typical fMRI studies only report significant effects, an underpowered statistical effect is much less likely to be detected when studies are combined meta-analytically. In other words, not only are small effects likely to be missed in individual studies, but if the data are not reported in individual studies as is typically the case, the combined sample sizes cannot be leveraged in meta-analyses. These real effects will be lost to meta-analyses. Moreover, when considering multiple papers in conjunction for determining a particular effect, Type II errors may conceptually overcorrect the literature if a finding is disregarded because follow-up studies are not permitted to report 'near significant' findings. It is important to remember that if one finds a $P = 0.05$ (corrected) effect, and then conducts an 'identical' replication (and assuming the effect size in study 1 is representative), there is only a 50% chance of significant replication (0.049 is in, 0.051 is not)!

WHAT SHOULD WE BE CORRECTING FOR?

It is conventional wisdom that it is more important for fMRI researchers to correct for multiple comparisons than for behavioral researchers to do so because fMRI researchers conduct so many more tests than behavioral researchers. Although there is no question that fMRI researchers conduct more tests than behavioral researchers in any given study, it is not clear why this is a criterion for who should be correcting and under what circumstances. How many behavioral science papers report dozens of statistical tests without any correction for multiple comparisons? If we are going to be serious about Type I errors, should not the number of tests reported in any paper be corrected for? While lip service is always given to $P < 0.05$ corrected for multiple comparisons, this in no way reflects the actual conventions of behavioral scientists. A randomly selected issue of *Journal of Personality and Social Psychology* (*JPSP*, August, 2000), a high-profile journal of the *American Psychological Association*, contained an average of 93 statistical tests per paper (range: 32–145 tests), excluding one paper that reported no statistical tests at all.[2]

If the goal of the push to avoid Type I errors in fMRI is to achieve a similar FDR as that observed in behavioral research, then we should be trying to match actual behavioral science research practices. As a first approximation, we used AlphaSim (Cox, 1996) to estimate the FDR for papers with 93 tests and then determined what cluster size, used in conjunction with an intensity threshold of $P < 0.005$ would produce the same FDR.[3] In the fMRI simulation, we assumed a $64 \times 64 \times 25$ matrix with a mask applied to include only voxels inside the brain (total voxels included: 39 838). We assumed voxel dimensions of 3.5 mm × 3.5 mm × 5 mm and a smoothing kernel of 6 mm full-width half-maximum. Based on one million simulations, we observed that $P < 0.005$ with a cluster size of 8 voxels achieves the same FDR as a 93 test study and a cluster size of 18 voxels achieves a FDR of 0.05.[4] We also compared the fMRI simulations to the *JPSP* paper with the least number of tests in the selected issue. To achieve the same FDR as a behavioral study with 32 tests, a *P*-value of 0.005 with a 9 voxel extent threshold is needed.

Whole-brain analyses using $P < 0.005$ with a 10 voxel extent threshold may not be equivalent to an FDR of 0.05 (though, $P < 0.005$ with a 20 voxel extent using the other scanning parameters we described does), but it is quite consistent with the FDR conventions used in actual behavioral research that fMRI researchers aim to emulate. If neuroimaging was already using inferential procedures as conservative with respect to Type I errors as actual practices in the behavioral sciences, why go further and further with methods that ensure an increasing number of Type II errors? We are not trying to suggest that $P < 0.005$ with a 10 voxel extent should be reified as a 'gold standard' criterion. The complexity of neuroimaging analyses suggests that a variety of standards might be appropriate in different contexts. We are focusing on this standard because it has been used so frequently in the past and is now being treated as an unacceptable criterion. In contrast, we think this is one of many different reasonable criteria for significance.

[2] We chose *JPSP* because this is one of the behavioral journals most relevant to social and affective neuroscientists. Looking at one journal is not meant to suggest this is the average or range for all journals, but we suspect the conclusions we draw will be relevant to other behavioral science journals as well. For instance, we looked at an issue of *Journal of Experimental Psychology: General* (November, 2009) and found that the mean number of statistical tests per paper was 49 (range: 15–82 tests).

[3] We are using FDR here in the generic sense of how frequently false alarms will be expected with a given set of parameters. We are not using the formal FDR methods as described by Genovese *et al*. (2002).

[4] It should be noted that others have run simulations suggesting that much larger cluster sizes are necessary (Vul *et al*., 2009; c.f. Lieberman *et al*, 2009), but these analyses included all of the voxels that are outside of the brain which are never actually tested. Thus, such analyses represent inaccurate assessments of the true FDR. It is also the case that these values will change with different voxel sizes and smoothing kernels, however the conceptual implications of these simulations remain the same.

One might respond that we should be correcting behavioral papers for the number of tests reported (i.e. we should change our statistical practices in the behavioral sciences to match Type I focus in the neuroimaging community). In that case, the articles in *JPSP* should have been using a per-test *P*-value threshold of 0.0005 and all that would have been left is the one article that did not report any tests. But if we are serious about correction, there is no logical reason to stop there. If the goal is to prevent Type I errors, perhaps each journal should require a correction for the number of tests reported in an entire issue of the journal ($P < 0.00005$). The selected issue of JPSP reported 932 tests; perhaps that should be the basis of correction because surely, some of those 932 tests will be significant as a result of chance alone. Perhaps there should be a correction for all the tests reported in a year of a journal ($P < 0.000005$) or perhaps for all the journals covering the same area (social psychology, emotion, memory) in a given month. This might be too difficult to decide on, so perhaps, instead of focusing on journals, we should focus on investigators and their labs. Perhaps a lab should have to correct for the total number of published results in a given year or maybe an investigator should have to continually update their correction factor based on the total number of results that they have published in their careers, with early career submissions having a more liberal correction factor and those in the National Academy of Science who have run countless tests, needing the most impressive results for them to qualify as significant (and, they would of course have to retract papers as their career progressed, finding that previous tests in old papers are no longer significant in light of their success and, ironically, contribution to the field).

The previous paragraph was meant to be hyperbolic in order to make an important point. As with just about everything else in our statistical analyses, corrections for multiple comparisons are about conventions and the conventions are arbitrary as long as they do not seriously offend our intuitions. There is no right way to correct for multiple comparisons that actually prevents Type I errors from being made. This is not how statistics operate and our attempt to treat statistics this way leads to serious underreporting of true effects that are likely to replicate. In any event, behavioral scientists have figured out a convention that works. They do not correct for multiple comparisons in actual practice, which assuredly produces Type I errors, but Type I errors that are not likely to replicate across multiple studies.

SOLUTIONS

Sample size

When money is no issue, increasing sample size is the surest within study method of reducing Types I and II errors. Unfortunately, with fMRI typically costing upwards of $500 per subject, money is very much an issue. This issue is likely to be exaggerated for new researchers and new research areas. We could limit fMRI research to a few elite labs, but this would likely undermine creativity and innovation in the field.

Replication and meta-analysis

The practical solution to addressing Types I and II errors within fMRI research is replication with systematic meta-analyses once a sufficient number of studies have been run. If the appropriate effects are reported in each paper, effects that would be too subtle to survive more conservative correction methods could still be examined in meta-analyses, which will help to prevent Type II errors. Similarly, if meta-analyses are taken seriously, Type I errors within individual papers cease to matter very much. False alarms will occur in individual studies no matter what precautions are taken, however, those false alarms should be randomly distributed spatially and thus not emerge as reliable effects in meta-analyses.

CONCLUDING SUMMARY

The press toward avoiding Type I errors is understandable. Humans desire certainty and conservative thresholding techniques help us to feel more certain that only real effects are being reported. But this analysis is flawed on two accounts: within-study *P*-values never guarantee that a false alarm has not occurred and thus there is no real certainty, but increasingly conservative thresholding techniques are absolutely certain to lead us to overlook real effects in the form of Type II errors.[5] Such effects are disproportionately likely to impact social and affective neuroscience studies and if such effects go unreported, meta-analyses will never have the data they need to find subtle effects that cannot be observed easily in individual studies. Moreover, FDR corrections end up being far more conservative than what is actually done in the behavioral sciences in which articles commonly report dozens of statistics at $P < 0.05$ uncorrected for multiple tests. We recommend placing a greater emphasis on replication and meta-analysis to determine which effects are real, and less emphasis on trying to determine the final truth from individual studies. With such an approach, Type I errors within individual studies matter far less because they are self-erasing, allowing for less conservative thresholding and fewer Type II errors.

[5] The companion piece by Bennett *et al.* (this issue) references a study in which a single dead salmon was the 'subject' in an fMRI scanner and active clusters were observed at $P < 0.005$ with a three voxel extent, pointing to a false alarm due to noise (rather than a true effect due to paranormal activity). This is used as a cautionary tale to argue in favor of using FDR correction techniques (which did not produce any false alarms). Given that FDR does allow for a 5% false alarm rate, FDR certainly could have produced a significant effect and yet had this happened, few would take that as an argument against FDR correction. More importantly, we think this example demonstrates the value of our own viewpoint quite clearly. They found a 3 voxel cluster in one fish, a Type I error no doubt. This same false alarm would not be present in the same location if 16 dead fish were all scanned and thus a group level analysis would not show this effect. And if an effect did emerge in a group analysis of 16 fish, it would not emerge in the same location in the next sample of 16 fish. Through data aggregation, false alarms are self-correcting but data that are never reported cannot be aggregated and thus Type II errors in fMRI are not self-correcting.

REFERENCES

Cox, R.W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–73.

Fisher, R.A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, *33*, 503–13.

Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., Noll, D.C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magnetic Resonance in Medicine*, *33*, 636–47.

Genovese, C.R., Lazar, N.A., Nichols, T.E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, *15*, 870–8.

Griffin, D.W., Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. *Advances in Experimental Social Psychology*, *24*, 319–59.

Lieberman, M.D., Berkman, E.T., Wager, T.D. (2009). Correlations in social neuroscience aren't voodoo: a reply to Vul et al. *Perspectives on Psychological Science*, *4*, 299–307.

Ochsner, K. (2007). How thinking controls feeling: a social cognitive neuroscience approach. In: Jones, E.H., Winkielman, P., editors. *Social Neuroscience: Integrating Biological and Psychological Explanations of Behavior*. New York, NY: Guilford Press, pp. 106–36.

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*, 274–90.

Wagner, A.D., Schacter, D.L., Rotte, M., et al. (1998). Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. *Science*, *281*, 1188–91.