A JOURNAL OF THE ASSOCIATION FOR PSYCHOLOGICAL SCIENCE aps

Big Correlations in Little Studies

Inflated fMRI Correlations Reflect Low Statistical Power—Commentary on Vul et al. (2009)

Tal Yarkoni

Washington University in St. Louis

ABSTRACT—*Vul, Harris, Winkielman, and Pashler (2009, this issue) argue that correlations in many cognitive neuroscience studies are grossly inflated due to a widespread tendency to use nonindependent analyses. In this article, I argue that Vul et al.'s primary conclusion is correct, but for different reasons than they suggest. I demonstrate that the primary cause of grossly inflated correlations in whole-brain fMRI analyses is not nonindependence, but the pernicious combination of small sample sizes and stringent alpha-correction levels. Far from defusing Vul et al.'s conclusions, the simulations presented suggest that the level of inflation may be even worse than Vul et al.'s empirical analysis would suggest.*

Vul, Harris, Winkielman, and Pashler (2009, this issue) argue that correlations in many cognitive neuroscience studies are grossly inflated due to a widespread tendency to use what they refer to as *nonindependent* analyses. A number of other commentators in this issue have taken issue with this conclusion, arguing either that nothing is wrong with the correlations fMRI studies have produced or that if anything is wrong, it's at least much less wrong than Vul et al. suppose. In this commentary, I adopt a different perspective. I argue that Vul et al.'s primary conclusion—that *r* values are inflated—is correct, but primarily for reasons other than those they suggest. Building on recent work by Yarkoni and Braver (in press), who discussed a number of conceptual and methodological issues related to the analysis of individual differences in fMRI studies, I demonstrate that the primary cause of inflated correlations in whole-brain fMRI analyses is the pernicious combination of small sample sizes and stringent alpha-correction levels. Far from defusing Vul et al.'s conclusions, the simulations presented suggest that the level of inflation may be even worse than Vul et al.'s empirical analysis would suggest.

Address correspondence to Tal Yarkoni, Campus Box 1125, Washington University in St. Louis, St. Louis, MO 63130; e-mail: tyarkoni@wustl.edu.

NONINDEPENDENT ANALYSIS IS NOT THE WHOLE STORY

Vul et al. suggest that many cognitive neuroscientists have used what they term nonindependent analyses to identify correlations: They first identify contiguous voxels that show a strong association between activation and behavior on the basis of surpassing some threshold and then conduct a second correlation test on the average of all voxels within the region, reporting only the latter *r* value as the final estimate of effect size. Vul et al. argue that this procedure capitalizes on chance and produces inflated *r* values by "selecting noise that exhibits the effect being searched for" (p. 279). Although this may be true to an extent, it can also be demonstrated that nonindependent testing isn't—and in fact, can't be—the source of all, or even much of, the inflation in *r* values.

To see this, suppose that we decide to test a correlational hypothesis using what Vul et al. would consider to be an independent analysis. We define 10 a priori regions of interest (ROIs) on the basis of some prior criterion (e.g., anatomy), average all the voxels within each region, and then correlate the mean level of activation within each region with behavior. We then report the resulting *r* value for all 10 ROIs in our published article. Are the resulting *r* values subject to inflation? The intuitive answer is no, because the procedure used to identify the ROIs is completely independent of the activation levels observed within those ROIs. But the truth is that the *r* values are only unbiased so long as we ignore any distinction between regions that show a significant effect and those that don't. When correlations are identified on the basis of attaining significance, they are indeed susceptible to inflation. Indeed, at sample size and *p* value parameters typical of fMRI studies, the degree of effect size inflation can potentially dwarf that suggested by Vul et al. (cf. their Fig. 5).

The presence of inflated correlations is readily demonstrated. Suppose that the actual population-level correlation in our hypothetical study is .4—an effect size that would be considered very large in most domains of psychology (cf. Meyer et al., 2001). Let's further suppose that there are 20 subjects in our sample

PERSPECTIVES on Psychological Science

and that we conduct each ROI-level test at an alpha-threshold of $p < .005$ ($p < .05$ corrected for 10 comparisons). A power calculation reveals that the probability of detecting a significant effect in each ROI is only 13%.[1] In other words, on average, only 1.3 of the 10 ROIs will show a significant effect in our sample. Critically, the mean *r* value within ROIs that do show a significant effect cannot possibly be .4, because the critical *r* value for a sample size of 20 tested at $p < .005$ is .6. So even if the population effect size is relatively large at .4, our hypothetical fMRI study will systematically inflate significant *r*s to a minimum of .6. On average, inflation will be still worse: Simulating 10,000 tests with the above parameters results in a mean significant *r* of .69. Clearly, gross inflation of *r* values can occur even in cases where researchers follow all of Vul et al.'s recommendations and use only independent analyses.

THE PROBLEM IS POWER

If inflation of *r* values still occurs independently of (non)independence, what can we attribute it to? The answer is statistical power—or, more accurately, a lack of power. A review of the vast literature on power is beyond the scope of this commentary (for accessible overviews, see Cohen, 1992; Maxwell, 2004; Sedlmeier & Gigerenzer, 1989); for present purposes it's enough to note that power is the probability of detecting a significant effect in one's sample given that that effect really exists in the population (i.e., the hypothesized association is nonzero).

Researchers underappreciate the fact that the power to detect between-subject effects is typically much lower than the power to detect within-subject effects of an equivalent magnitude. Yarkoni and Braver (in press) plotted power curves for correlation tests and *t* tests across a range of sample sizes and alpha levels commonly used in fMRI studies (reproduced here as Fig. 1). We showed, for example, that the power to detect a canonically large effect size of $d = 0.8$ (Cohen, 1988) using a one-sample *t* test in a sample of 20 subjects tested at $p < .001$ is approximately 40%. In contrast, the power to detect a roughly equivalent *r* of approximately 0.36 in the same sample is only 2.6%. The importance of this point is difficult to overstate: Under reasonable assumptions, the power to detect correlational effects may be as little as 5%–10% of the power to detect similar-sized within-subject effects. And testing at a more liberal level of $p < .05$ doesn't help much: In that case, the power discrepancy is 92% versus 32%.

Yarkoni and Braver (in press) reviewed a number of negative consequences associated with the use of low-powered correlational tests in fMRI studies. The most obvious is the failure to detect real effects—that is, Type II error, which is simply the complement of power. Needless to say, failing to detect real effects is never a good thing, and investigators should strive to always conduct adequately powered studies. However, a less widely appreciated consequence of low power is the aforementioned inflation of significant effect sizes. This point is illustrated systematically in Figure 2, which demonstrates that for all but the largest fMRI sample sizes (i.e., anything up to at least 30 subjects), one can expect massive inflation of significant *r* values for all but the strongest population effects. For example, a population effect of 0.3 will show up, on average, as a significant *r* of 0.73 when identified in a sample of 20 subjects tested at a $p < .001$ threshold. In fact, the mean significant *r* value for 20 subjects at $p < .001$ shows almost no movement as a function of the real correlation size, simply because the critical *r* value is already so high (.65).

The combination of low power and effect size inflation can easily lead to misinterpretation of fMRI results if investigators are not careful. Many behavioral measures probably correlate relatively diffusely with brain activation in the general population. Suppose, for example, that there is a .3 correlation between a broad personality dimension like neuroticism and activation in half of the brain when people look at aversive pictures. An investigator who conducts a whole-brain analysis in a sample of 20 subjects is unlikely to detect more than a couple of relatively circumscribed neuroticism-related regions due to low power; moreover, correlations will be grossly inflated within the identified regions, hovering around 0.75–0.80 on average. So although the correct characterization may be that a given brain–behavior relationship is modest in size and spatially diffuse, a small-sample whole-brain analysis is likely to instead conclude that it's extremely strong and highly selective.

JUST HOW STRONG ARE THESE CORRELATIONS?

There is, admittedly, a very large assumption underlying the pessimism suggested by the above discussion. Namely, one has to assume that the real size of brain–behavior correlations is approximately the same as, or not much larger than, the correlations typically observed in behavioral studies. One might think this is pretty unlikely, because empirically, brain–behavior correlations appear to be huge. But such reasoning is circular: The primary reason people suppose that brain–behavior correlations are so much stronger than behavior–behavior correlations is because of the very same effects that Figure 2 calls into question. So instead, we need to turn to other lines of evidence and argument. Here, I'll focus on just three reasons to think that population correlations between brain activation and behavior aren't really as big as previous results suggest.

First, there's the admittedly subjective argument from plausibility. It is worth considering what exactly investigators are claiming when they report an *r* of, say, 0.85. The implication is that over 70% of the variance in a dimension like neuroticism or

[1]For the sake of simplicity, I assume throughout this article that all tests are conducted on independent observations. The numbers change slightly in the presence of nonindependence, but the central point remains unchanged. Similarly, I deal only with intensity (voxel-wise) thresholds and ignore cluster thresholds, which will reduce power further when combined with a constant intensity threshold.

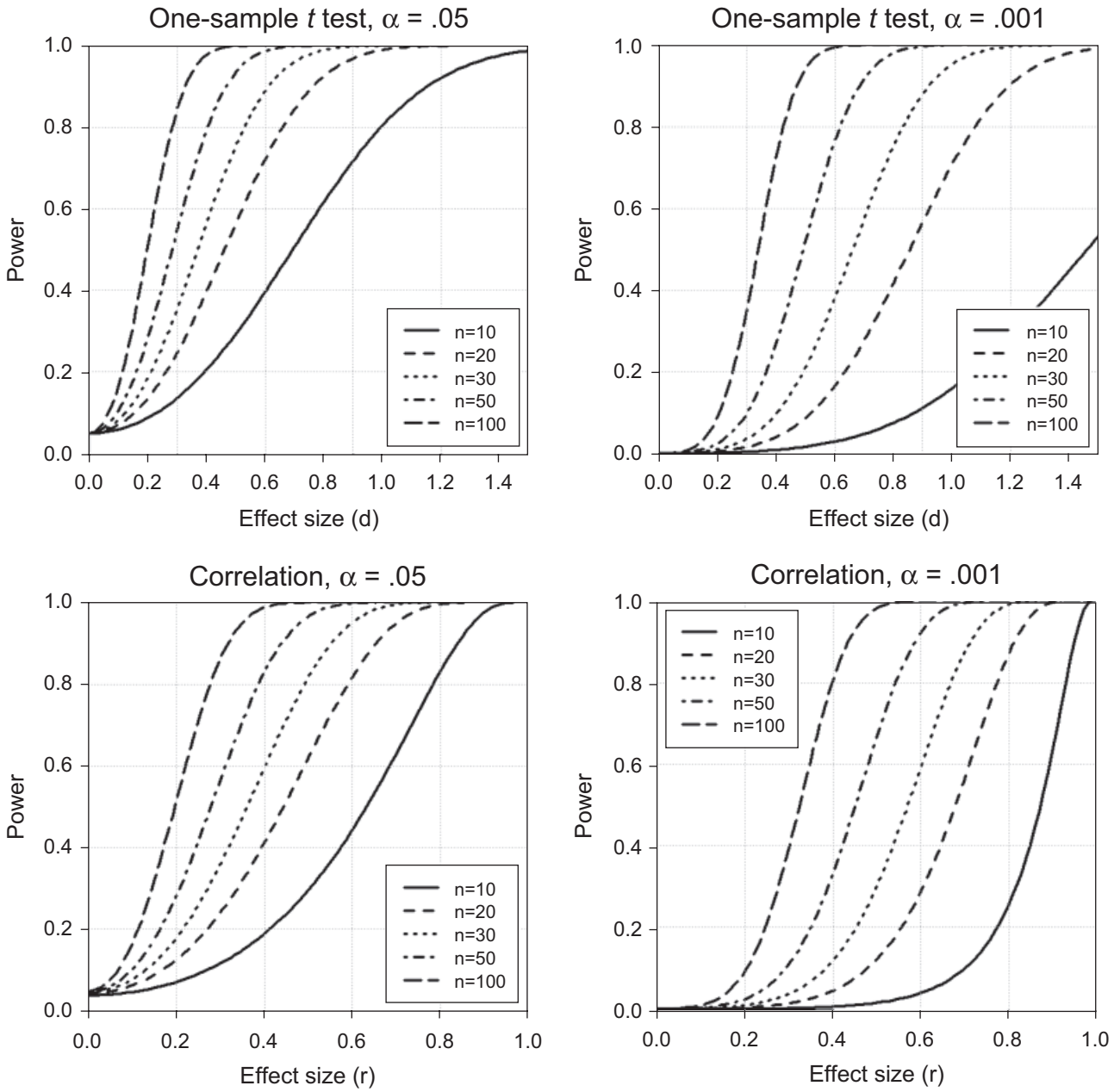

Fig. 1. Statistical power as a function of test type (top: one-sample *t* test; bottom: Pearson's *r*), alpha level (left: $p < .05$; right: $p < .001$), sample size, and effect size.

empathy or fluid intelligence is explained by activation in just one brain region for a very specific contrast in a task that is probably not all that reliable to begin with (both Vul et al. and Yarkoni & Braver, in press, review a number of studies that, collectively, raise questions about the reliability of fMRI). In systematic studies of psychological and biomedical effect sizes (e.g., Meyer et al., 2001), one rarely encounters correlations greater than .4. Correlations of .85 are practically unheard of, unless they are trivial (e.g., between two self-report measures of the same construct). So investigators should be very careful when concluding that the huge correlations routinely observed in fMRI studies provide accurate estimates of population effect sizes.

Second, an investigator who believes in big *r*s has to explain why it is that most within-subject contrasts identify relatively distributed patterns of activation, whereas correlational analyses do not. Why is it that we tend to see many more "selective" effects in correlational analyses? There's no good conceptual reason to suppose that individual differences effects are so much more localized than within-subject effects. But that discrepancy is exactly what one would expect if power is substantially lower for between-subject analyses than for within-subject analyses. An underpowered fMRI analysis will consistently produce spatially circumscribed and numerically inflated effects.

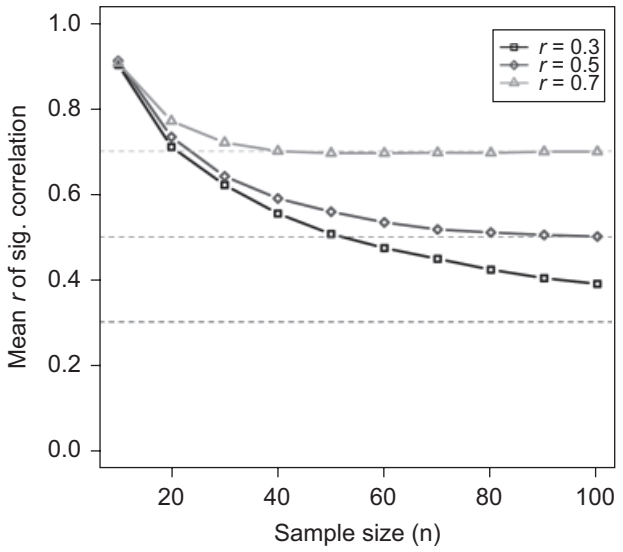

Fig. 2. Inflation of significant *r* values as a function of sample size (*x* axis) and population effect size (lines). Each point represents the result of 10,000 simulated correlation tests, each conducted at a threshold of $p < .001$, reflecting the most commonly used whole-brain threshold. Dashed lines represent the true correlation size; solid lines represent the mean observed correlation in the sample for only those tests that produce significant results.

Third, a believer in big *r*s needs to explain why big *r*s appear to be the exclusive province of small-sample studies. Vul et al.'s meta-analysis demonstrates that correlations >.8 are not rare in fMRI (see their Fig. 5). But virtually all of the data points come from studies with sample sizes < 30, and the majority of them come from studies with samples far smaller than that. To my knowledge, no fMRI study with 50 or more subjects has ever reported a correlation of .8 or greater. If one believes that brain–behavior correlations of that strength exist in the population, the absence of any large-sample studies reporting such correlations is inexplicable. As sample size grows, the variance around the effect size estimate decreases, providing an increasingly better estimate of the population effect. So why wouldn't we see larger *r* values in big studies?

In fact, what tends to happen is exactly the opposite: As sample size grows, effects shrink. Yarkoni and Braver (in press) gave the example of two recent studies by Gray and colleagues (Gray & Braver, 2002; Gray et al., 2005). The first study ($N = 14$) identified very strong correlations (*r*s ranging from −.63 to −.84 across different conditions) between behavioral activation sensitivity (a personality dimension conceptually related to extraversion) and anterior cingulate cortex activation during a working memory task (Gray & Braver, 2002). The second study ($N = 53$) replicated the effect, but with a much smaller correlation of −.28 (Gray et al., 2005). As a study with 53 participants provides much more reliable effect size estimates than one with 14 participants, the likeliest explanation for the discrepancy is that the large *r*s in the first study were grossly inflated by the aforementioned combination of small sample size and stringent alpha thresholds. Indeed, if you suppose that the real population correlation was in the neighborhood of, say, .3, then even a study with 53 subjects testing at $p < .05$ would only have a 59% chance to detect the effect. So, if anything, Gray et al. may have been lucky to replicate the effect in their second study. Needless to say, the idea that you might need 50 subjects just to have approximately even odds of detecting an effect within an a priori ROI tested at $p < .05$ may be a difficult one to swallow, but it's far likelier to hold true than the notion that one can get by with only 12 or 15 subjects.

IMPLICATIONS

Truth be told, it is hard to quantify exactly how much of a problem lack of power is for correlational analyses in fMRI studies, because we don't conclusively know what type of effect sizes exist in the population. But the above considerations suggest that it is very unlikely that most brain–behavior correlations are stronger than, say, .5—an effect size that would already seem extremely large to most behavioral researchers. An fMRI study with 20 subjects would have a 61% chance to detect a .5 correlation even at a "liberal" threshold of $p < .05$—not terrible, but certainly not adequate. But if an investigator decides to conduct a whole-brain analysis at, say, $p < .001$, power drops to just 12%. To achieve a conventionally acceptable level of 80% power, it would take 29 subjects at $p < .05$ and a full 60 at $p < .001$.

The implication is that it is almost certain that the vast majority of whole-brain correlational analyses (a) identify only a fraction of the effects that really exist in the population, (b) grossly inflate the apparent size of those effects that researchers are lucky enough to detect, and (c) promote a deceptive illusion of highly selective activation. Far from dispelling Vul et al.'s conclusions, these considerations suggest that matters may be even worse than Vul et al. suggest. Cognitive neuroscientists don't just have to worry about the inflated *r* values that they do see, they also need to worry about the many correlations that aren't detected due to insufficient power. Consistently running studies that are closer to 0% power than to 80% power is a sure way to ensure a perpetual state of mixed findings and replication failures.

WHAT TO DO ABOUT IT

Vul et al.'s chief recommendation is that investigators always use independent analyses. Although this recommendation will help reduce inflation of correlations to some extent, it doesn't address the underlying problem of insufficient power. Yarkoni and Braver (in press) made a number of suggestions for dealing with low power and inflated *r* values. The most obvious, but also most painful, solution is to increase sample size. Simply put, if the primary objective of a study is to detect individual differences in brain activation, a sample size of 20 should be considered flatly unacceptable. One would need the population correlation to be .6 in order to have an 80% chance of detecting the effect in the sample at $p < .05$; that's a gamble one should be very hesitant to take. If one intends to conduct whole-brain analyses, a more

reasonable sample size might be 50—and that would still provide only a 66% chance to detect correlations of .5 at $p < .001$! Needless to say, fMRI samples this large are extremely expensive and time-consuming to collect. But the fact that it's difficult to collect large samples is not a good enough reason to keep running underpowered studies. Most cognitive neuroscientists would be skeptical of an investigator who planned to conduct a within-subject study with only 7 subjects without first doing a power analysis; yet, for many combinations of effect size and *p* value, a one-sample *t* test with 7 subjects actually provides more power than a correlation test with 20 subjects (cf. Fig. 1).

Second, investigators should perform power calculations prior to beginning fMRI data collection and should generally report those calculations in their manuscripts. Reviewers and editors should similarly be encouraged to request or require authors to report power calculations when none are provided. Many power analysis tools are freely available either as stand-alone applications or as add-ons to popular statistics packages,[2] and many Web sites provide instantaneous power calculations for various statistical procedures. The time investment required to perform a series of power calculations is negligible, and the potential pitfalls of failing to do so enormous, so investigators have every incentive to be diligent about power considerations.

Third, with respect to inflation of significant *r* values, investigators should either pay little or no attention to the size of correlations or report all correlations with confidence intervals and pointedly emphasize their likely unreliability. The former measure seems excessively strong until one remembers that almost nobody ever reports effect sizes for *t* tests or analyses of variance, despite many journals' explicit encouragement to do so. Ideally, psychologists would focus on those measures of effect size that stem from much more powerful within-subject tests and would question or even ignore those that come from lower powered correlational tests. In practice, however, we seem to do exactly the opposite. So outright elimination of *r* from the pages of our journals should be considered a viable option, if only for consistency's sake. The alternative—reporting confidence intervals around every *r*—is probably better, but is much more cumbersome.

Finally, reviewers should be skeptical of any correlational study that purports to find a "selective" relation between brain and behavior. Unless an fMRI study has an extremely large sample size, investigators should be very wary of claiming that some regions show an effect whereas others do not. Single and double dissociations are enormously powerful tools when used in the context of a high-powered within-subjects study, but it is difficult to conceive of many situations in which a correlational study would have enough power to make a corresponding claim. Suppose, for example, that two a priori anatomical ROIs are tested in a sample of 20 subjects and that Region A is found to correlate significantly with Measure X but not Y whereas Region B correlates significantly with Measure Y but not X. If the real correlation between A and X is 0.5, and the real correlation between A and Y is 0.3, there is a 62% chance to detect the A–X correlation, but only a 24% chance to detect the A–Y correlation. Clearly, single and even double dissociations will not be hard to come by when power is low.

[2]I used the free software package *R* (*R* Foundation for Statistical Computing, Vienna, Austria) and the *pwr* library add-on to perform all of the calculations and simulations reported here.

CONCLUSIONS

In sum, the present considerations suggest that Vul et al. are essentially correct with respect to their primary conclusion: Correlations in cognitive neuroscience are inflated, and probably to an even greater extent than Vul et al. suggest. However, this inflation primarily reflects a lack of power rather than the use of nonindependent analyses. The bad news is that many correlational effects that seemed too good to be true almost certainly were too good to be true; *r* values on the order of .7 to .8 probably shouldn't be trusted. What's worse, for every significant *r* that made it to print, there were probably many others that were overlooked and that now sit patiently in people's brains waiting to be discovered by higher powered studies. Admittedly, that's pretty bad news. But the good news is that there's no mystery behind the inflation of *r*; we know exactly what to do about low power, and it's just a matter of spending the time and money to do it.

REFERENCES

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.

Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological Bulletin*, *112*, 155–159.

Gray, J.R., & Braver, T.S. (2002). Personality predicts working-memory-related activation in the caudal anterior cingulate cortex. *Cognitive, Affective, & Behavioral Neuroscience*, *2*, 64–75.

Gray, J.R., Burgess, G.C., Schaefer, A., Yarkoni, T., Larsen, R.J., & Braver, T.S. (2005). Affective personality differences in neural processing efficiency confirmed using fMRI. *Cognitive, Affective, & Behavioral Neuroscience*, *5*, 182–190.

Maxwell, S.E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L., Dies, R.R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128–165.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies. *Psychological Bulletin*, *105*, 309–316.

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*, 274–290.

Yarkoni, T., & Braver, T.S. (in press). Cognitive neuroscience approaches to individual differences in working memory and executive control: Conceptual and methodological issues. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of individual differences in cognition: Attention, memory and executive control*. New York: Springer.