Chapter 6
Cognitive Neuroscience Approaches to Individual Differences in Working Memory and Executive Control: Conceptual and Methodological Issues

Tal Yarkoni and Todd S. Braver

Analyses of individual differences play an important role in cognitive neuroscience studies of working memory and executive control (WM/EC). Many studies examining the neural substrates of working memory have relied upon correlations between brain activity and either task performance measures or trait measures of cognitive ability. However, there are important conceptual and methodological issues that surround the use of individual difference measures to explain brain activation patterns. These issues make the interpretation of correlations a more complex endeavor than is typically appreciated.

In this chapter, we review several issues that have been of long-standing concern in behavioral research on individual differences and that are equally relevant to cognitive neuroscience studies of executive control and working memory. The chapter is structured into three parts. In the first part, we provide a selective review of the literature in this domain, highlighting the most common individual difference approaches, as well as emerging trends. The scope of the review is restricted to human neuroimaging studies using fMRI methods, since this is the domain in which most of the relevant work has been conducted. However, we expect that many of the issues and relevant findings will apply equally well to other cognitive neuroscience methods (e.g., PET, ERP, TMS, etc).

The second part discusses the conceptual relationship between within-subject and individual differences analyses, focusing particular attention on situations in which within-subject and individual differences analyses produce seemingly discrepant results. Finally, in the third part, we selectively review a number of statistical and methodological concerns that arise when conducting individual differences analyses of fMRI data, including the relative lack of statistical power, the absence of data concerning the reliability of individual differences in the BOLD signal, and the deleterious effects of outliers.

An Overview of Individual Differences Approaches

Because of the high cost of conducting fMRI research and the importance of using large samples in individual differences research (see "Methodological and Statistical Considerations"), individual differences analyses in fMRI studies of WM/EC are typically conducted as an opportunistic complement to within-subject analyses and are rarely the primary focus of a study. As such, most individual differences analyses of fMRI data are subject to many or all of the following constraints:

T. Yarkoni and T.S. Braver (✉)
Departments of Psychology & Radiology, Washington University, Campus Box 1125, St. Louis, MO 63130, USA
e-mail: tbraver@wustl.edu

A. Gruszka et al. (eds.), *Handbook of Individual Differences in Cognition: Attention, Memory, and Executive Control*, The Springer Series on Human Exceptionality,
DOI 10.1007/978-1-4419-1210-7_6, © Springer Science+Business Media, LLC 2010

1. Relatively small sample sizes (the current norm appears to be between 15 and 20 participants);
2. Systematic exclusion of participants with certain characteristics in order to ensure relatively homogeneous samples (e.g., screening participants based on age, gender, or WM capacity);
3. Use of experimental manipulations that are chosen in part because they are known to produce consistent changes in behavior across participants;
4. Titration of task difficulty to ensure that all participants' performance levels fall within bounds amenable to within-subject analyses (e.g., a minimum cutoff of 80% accuracy);
5. Little or no measurement of preexisting differences in participants' cognitive abilities and/or personalities (e.g., as might be assessed using batteries of psychometric measures).

All of these constraints have a deleterious effect on the probability of detecting individual differences effects and/or generalizing significant effects to the broader population. While practical considerations make it difficult to overcome many of these limitations (e.g., collecting data from much larger samples is often not viable, from a financial perspective), a number of different approaches have been used to increase detection power and/or support stronger inferences when analyzing individual differences. We selectively highlight a number of fMRI studies that have used such approaches.

Continuous vs. Extreme Groups Designs

While participant samples in most fMRI studies are randomly sampled (subject to general constraints on demographic variables such as age and gender), some studies have attempted to maximize power to detect individual differences effects by using *extreme groups* (EG) designs in which participants are stratified into two or more groups based on their scores on some variable of interest (e.g., WM capacity or fluid intelligence; Larson, Haier, LaCasse, & Hazen, 1995; Lee et al., 2006; Mecklinger, Weber, Gunter, & Engle, 2003; Osaka et al., 2003). This approach increases the power to detect effects by reducing the variance between participant groups relative to the variance within groups, thereby inflating effect sizes and making them easier to detect.[1] For example, Lee and colleagues selected participants into two groups based on their scores on the Raven's Advanced Progressive Matrices (RAPM; Raven, Raven, & Court, 1998), a putative measure of *g*, or general intelligence (Lee et al., 2006). Both groups showed increased activation in frontal and parietal regions when performing a high *g*-loaded task relative to a low *g*-loaded cognitive task; however, the increase was substantially greater for the high-*g* group than the low-*g* group. The effect sizes observed (with peak correlations of about 0.8 between RAPM scores and brain activation) would almost certainly have been smaller if a random sample of participants had been recruited.

The increase in power obtained using EG designs is not without its costs (for review, see Preacher et al., 2005). Inflated effect sizes produced by EG designs may lead researchers to overestimate the importance of the effects in the general population (one simple solution to this problem is to pay less attention to effect sizes that result from EG studies). Moreover, an EG design will preclude identification of nonlinear effects (e.g., a curvilinear relationship between brain activation and task performance) that require examination of a full distribution of scores. Nevertheless, in cases where researchers are interested in a specific dimension of individual differences, have limited data collection resources, and detection power is more important than accurate characterization of effect size, EG samples may be preferable to random samples.

[1] It should be noted that extreme groups designs are *not* equivalent to post-hoc dichotomization of participants based on a median split of scores on some variable of interest. The latter approach substantially *reduces* power and is almost never justified (Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002; Preacher, Rucker, MacCallum, & Nicewander, 2005). The same is true for subsampling extreme groups from a larger sample (e.g., performing a *t*-test comparing the lowest-scoring ten participants to the highest-scoring ten participants drawn from a full sample of 60 participants).

Behavioral Approaches

As noted above, the variables used to predict individual differences in brain activation are selected opportunistically in most fMRI studies of WM/EC. Most commonly, they include simple performance measures of in-scanner behavioral performance such as response accuracy (Callicott et al., 1999; Gray, Chabris, & Braver, 2003; Yarkoni, Gray, & Braver, submitted) or reaction time (Rypma, Berger, & D'Esposito, 2002; Rypma & D'Esposito, 1999; Schaefer et al., 2006; Wager, Sylvester, et al., 2005). This approach is straightforward to implement (typically requiring no special consideration when designing an experiment) and can provide valuable insights into the relationship between task performance and activation increases or decreases. On the other hand, it does not allow researchers to differentiate *state* effects (some participants may perform better or worse during a given session for relatively uninteresting reasons) from *trait* effects (some participants consistently perform better than others). Moreover, the source of a particular brain–behavior correlation may be relatively unclear (e.g., is a positive relationship between dorsolateral prefrontal cortex (DLPFC) and verbal WM performance evidence of increased WM capacity (a trait effect) or of greater effort expenditure (a state effect)?).

To overcome such limitations, a small number of studies have related brain activation not only to in-scanner behavioral tasks but also to tasks or measures administered outside of the scanner. These include both standard ability-based measures of WM span or fluid intelligence (e.g., Geake & Hansen, 2005; Gray et al., 2003; Haier, White, & Alkire, 2003; Lee et al., 2006) and questionnaire-based personality measures of dimensions such as extraversion and neuroticism (Gray & Braver, 2002; Gray et al., 2005; Kumari, Ffytche, Williams, & Gray, 2004). For example, Yarkoni and colleagues recently used multiple psychometric measures of cognitive ability to demonstrate that activation in a region of medial posterior parietal cortex (PPC) during a 3-back WM task correlated significantly with fluid and spatial abilities but not with crystallized or verbal abilities (Yarkoni, Gray, et al., submitted). This finding suggested that individuals who recruited mPPC activation more extensively (and performed more accurately) may have implicitly relied on a spatial strategy, despite the fact that the task itself had no overt spatial element. Such an inference would not be possible solely on the basis of correlations between brain activation and in-scanner performance, and requires the use of additional behavioral measures.

Another compelling illustration of the utility of behavioral measures is found in studies investigating the relationship between behaviorally measured WM capacity and the influence of dopaminergic drugs on in-scanner task performance and brain activation (S. E. Gibbs & D'Esposito, 2005; S. E. B. Gibbs & D'Esposito, 2006; Kimberg, Aguirre, Lease, & D'Esposito, 2001; Mattay et al., 2000; Mehta et al., 2000). At the group level, such studies have produced somewhat mixed results: administration of a dopamine-enhancing agent may improve WM performance and/or decrease cortical activation (Kimberg et al., 2001; Mehta et al., 2000), produce no main effect (S. E. Gibbs & D'Esposito, 2005), or even *impair* performance (S. E. B. Gibbs & D'Esposito, 2006). Importantly, however, these mixed effects are moderated by individual differences in baseline WM capacity. Across several studies, low-capacity individuals consistently benefit more from dopaminergic drugs than high-capacity individuals (S. E. Gibbs & D'Esposito, 2005; Kimberg, D'Esposito, & Farah, 1997; Mattay et al., 2000; Mehta et al., 2000), a finding that sheds considerable light on what would otherwise be a murky literature, and may have important practical implications for the use of such drugs.

Statistical Approaches

Most individual differences analyses in fMRI studies consist of basic parametric or nonparametric correlation tests (e.g., Pearson's and Spearman's correlation coefficients, respectively). However, some studies have gone beyond simple correlational analyses and have applied more sophisticated analytical procedures common in psychometric studies of WM/EC to fMRI data. One such approach

is statistical mediation analysis, which indicates whether a set of correlations between variables A, B, and C is consistent with a causal model postulating that the effect of variable A on variable C is at least partly explained through the mediating role of variable B (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). Several fMRI studies have used mediation analyses to identify brain regions that significantly mediate the relationship between two behavioral variables (e.g., Richeson et al., 2003; Tom, Fox, Trepel, & Poldrack, 2007; Yarkoni, Braver, Gray, & Green, 2005), though the number of applications to the domain of WM/EC remains limited. One study notable for its use of mediation analysis in a relatively large sample ($n=50$) was reported by Gray and colleagues (Gray et al., 2003). They conducted an fMRI experiment in which participants completed a standard measure of fluid intelligence (RAPM) and subsequently performed a challenging 3-back WM task in the scanner. The study's design enabled the authors to identify a region in the left lateral prefrontal cortex (PFC) that statistically *mediated* the correlation between fluid intelligence and 3-back response accuracy – a stronger inference than would be afforded simply by observing a correlation between brain activation and task performance.

Another statistical technique widely used in psychometric WM/EC research is structural equation modeling (SEM; Bollen, 1989; Kline, 1998), which enables researchers to estimate, evaluate, and develop causal models of the relationships between variables (Friedman & Miyake, 2004; Kane et al., 2004). To date, cognitive neuroscientists interested in WM/EC have used SEM primarily to model interactions between different brain regions and/or WM task conditions (for review, see Schlösser, Wagner, & Sauer, 2006), and only secondarily to relate individual differences measures of WM to brain activation (but see, e.g., Glabus et al., 2003; Kondo et al., 2004). However, in principle, a unified SEM framework could integrate both behavioral and neural data – e.g., by mapping latent variables derived from behavioral measures onto distinct brain networks (Kim, Zhu, Chang, Bentler, & Ernst, 2007). Of course, such efforts are likely to be hampered by considerable practical obstacles – e.g., the need for large sample sizes, multiple in-scanner sessions, etc. – so SEM approaches on the scale of existing behavioral analyses (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, Tuholski, Laughlin, & Conway, 1999; Kane et al., 2004) may not be viable.

An alternative multivariate technique used in a number of WM/EC studies involving individual differences (e.g., Caplan, McIntosh, & De Rosa, 2007; Della-Maggiore et al., 2000; Grady et al., 1998) is Partial Least Squares (PLS; McIntosh, Bookstein, Haxby, & Grady, 1996; McIntosh & Lobaugh, 2004). Like SEM, PLS focuses on network-level activation rather than individual voxels or brain regions; however, unlike other multivariate approaches, PLS seeks to identify patterns of brain activation that covary maximally with a reference set of variables (e.g., an experimental design matrix or a set of individual difference variables). For example, Caplan and colleagues recently used PLS to identify two distributed brain networks that independently predicted individual differences in the successful resolution of proactive interference (Caplan et al., 2007). The use of a multivariate approach allowed the authors to model brain activation at the level of functional networks rather than isolated regions, an approach that would not have been possible using conventional univariate methods. However, PLS remains susceptible to many of the limitations of SEM (e.g., the fundamental need for a large sample size when focusing primarily on individual differences).

The Relationship Between Within-Subject and Individual Differences Analyses

How are individual differences analyses conceptually related to more common within-subject analyses (e.g., paired *t*-tests)? One common intuition is that the results of the two types of analyses should tend to converge. That is, regions in which an experimental manipulation elicits greater activation on a within-subject level should also show reliable between-subjects variation, such that

activation is greater in individuals who capably perform the task than in individuals who do not. There is some empirical support for this idea; for example, DLPFC activation reliably increases as a function of WM load (Braver et al., 1997; Callicott et al., 1999), and several studies have found a positive correlation between DLPFC activation and higher WM capacity or greater fluid intelligence (Gray et al., 2003; Lee et al., 2006). On the other hand, there are arguably more instances in the literature of within-subject and between-subjects analyses producing effects in conceptually *opposing* directions. For example, older adults often show greater prefrontal activation than younger adults when performing effortful cognitive tasks despite exhibiting a poorer level of behavioral performance (Cabeza, Anderson, Locantore, & McIntosh, 2002; Reuter-Lorenz, 2002). In young adults, many studies have similarly found greater activation in regions such as lateral PFC and anterior cingulate cortex (ACC) in participants who perform poorly than those who perform well. The ACC is thought to be involved in monitoring for conflict, and in several studies, individuals who perform more poorly on tasks involving cognitive conflict have shown greater ACC activation (Bunge, Ochsner, Desmond, Glover, & Gabrieli, 2001; Hester, Fassbender, & Garavan, 2004; MacDonald, Cohen, Stenger, & Carter, 2000; Wager, Sylvester, et al., 2005). Left ventrolateral PFC (VLPFC) is thought to be involved in resolving proactive interference (Jonides & Nee, 2006), yet individuals who are good at overcoming proactive interference show *less* VLPFC activation than individuals who are not (Nee, Jonides, & Berman, 2007).

There is often a plausible explanation for the opposing directions of within-subject and individual differences effects. For example, the observation that susceptibility to interference during cognitive conflict-inducing tasks correlates positively with ACC activation might be attributed to the fact that the ACC tracks the amount of *input* conflict rather than the degree to which its resolution is successful (e.g., Wager, Sylvester, et al., 2005). More generally, frontoparietal hyperactivations in older adults and poorly performing young adults are often attributed to compensatory processing or differences in *neural efficiency* (Gray et al., 2005; Haier et al., 1992; Larson et al., 1995; Rypma et al., 2006). That is, older adults and inefficient young adults are thought to require greater effort expenditure to achieve the same level of performance as highly performing young adults, leading to relative increases in frontoparietal activation. This interpretation affords easy reconciliation of the apparent contradiction between individual differences and within-subject effects. However, the explanatory power of this perspective is limited, as it provides no indication as to *why* individuals might vary considerably in cognitive efficiency, or why correlations between frontoparietal activation and cognitive performance are positive in some studies but negative in others.

An additional consideration is that correlations between mean reaction time (RT), a primary measure of behavioral performance, and brain activation may simply reflect basic properties of the blood oxygen level dependent (BOLD) signal detected by fMRI. In a recent multistudy analysis, we demonstrated that trial-by-trial differences in reaction time correlate positively with lateral PFC and MFC activation in a task-independent manner (Yarkoni, Barch, Gray, Conturo, & Braver, submitted). The explanation accorded to this finding was that frontal increases on slower trials reflect linear summation of the hemodynamic response over time (cf. Burock, Buckner, Woldorff, Rosen, & Dale, 1998; Dale & Buckner, 1997). While we did not report individual differences analyses of RT, a temporal summation account should hold both within and across individuals. That is, individuals who sustain attention to task-relevant information for a longer period of time may show a greater summation of the BOLD signal in frontal regions, irrespective of whether that time is used efficiently or not. Given this possibility, caution should probably be exercised when using individual differences in RT as a predictor of brain activity. A preferable approach is to rely on response accuracy as a measure of performance while statistically controlling for individual differences in RT. Unfortunately, this approach is not viable for many tasks in which RT is the primary measure of performance. Nevertheless, given that it is possible to include RT as a trial-by-trial covariate in fMRI analyses, effectively controlling for this source of variance, it might be useful in such studies to test how the inclusion of an RT covariate influences individual difference effects.

Spatial Dissociations

Discrepancies between within-subject and individual differences analyses need not manifest as conceptually opposing effects within the same brain regions. Often, individual differences analyses simply fail to reveal *any* significant effects in regions that show a robust within-subject effect. In such cases, lack of statistical power, which we discuss at length in the next section, is a likely culprit, because individual differences analyses almost invariably have considerably lower power than within-subject effects. Alternatively, individual differences analyses and within-subject analyses may both reveal significant effects but in spatially dissociable regions (Bunge et al., 2001; Locke & Braver, 2008; Yarkoni, Gray, et al., submitted). For example, in a recent large-sample ($n = 94$) fMRI study using a 3-back WM task, we found a dissociation between frontoparietal regions canonically implicated in cognitive control and a region of medial posterior parietal cortex (mPPC) not usually implicated in WM/EC (Yarkoni, Gray, et al., submitted). Frontoparietal regions showed strong within-subject effects of trial difficulty but inconsistent associations with response accuracy either within or across individuals. In contrast, mPPC activation showed relatively weak effects of trial difficulty but robust associations with response accuracy both within and across subjects.

How should spatial dissociations between within-subject and individual differences effects be interpreted? On the one hand, such findings may seem counterintuitive, because regions that do not appear to be recruited *on average* during performance of a task may seem like unlikely candidates for individual differences effects. Indeed, some evidence suggests that the BOLD signal tends to be more reliable across individuals in regions that show strong within-subject effects than in those that do not (Aron, Gluck, & Poldrack, 2006; Specht, Willmes, Shah, & Jaencke, 2003). On the other hand, there is no logical necessity for within-subject and between-subject sources of variance to produce converging results. Indeed, from a statistical standpoint, within-subject and between-subject effects are in tension, because between-subjects variance counts as error in within-subject analyses, and vice versa. This observation raises concerns that one of the most common individual differences approaches in fMRI studies – namely identifying regions-of-interest (ROIs) on the basis of within-subject analyses and subsequently testing them for correlational effects – may actually *reduce* the probability of detecting significant effects. Omura and colleagues recently advocated precisely the opposite approach, suggesting that researchers interested in correlational effects should focus their search on regions in which between-subjects variance is largest relative to within-subject variance (Omura, Aron, & Canli, 2005). However, the viability of the latter approach depends on the assumption that the signal in regions with a high BS/WS variance ratio is reliable. If, as the studies cited above suggest, reliability tends to be highest in regions that show the strongest *within*-subject effects, there would be little utility in such an approach. Additional empirical studies are needed in order clarify the issue.

Integrative Interpretation of Within-Subject and Individual Differences Results

Although directional or spatial dissociations between within-subject and individual differences analyses may be difficult to interpret, they can potentially also serve as a powerful tool for characterizing the functional role of different brain regions. Within-subject and individual differences analyses reflect independent sources of variance, are sensitive to different kinds of processes, and afford qualitatively different conclusions. We focus here on two kinds of inferences afforded when combining the two forms of analysis. First, within-subject analyses are, by definition, most sensitive to processes that are consistent across all participants, including those that are *necessary* for performing a task. For example, the fact that the "task-positive" frontoparietal network (Fox et al., 2005) is activated

during virtually all tasks involving cognitive effort (Buchsbaum, Greer, Chang, & Berman, 2005; Duncan & Owen, 2000; Owen, McMillan, Laird, & Bullmore, 2005; Wager, Jonides, & Reading, 2004) likely indicates that an intact frontoparietal network is necessary in order to maintain a minimal level of goal-directed attention; however, this network may play little or no role in supporting many of the task-specific processes that distinguish one effortful cognitive task from another. In contrast, variability in performance across individuals is most likely to reflect those processes that can be recruited in varying degrees to produce incremental gains in performance. These include both processes that can vary quantitatively in strength (e.g., the amount of effort one exerts during the task) as well as qualitatively different strategies that may vary across individuals (e.g., in an *N*-back task, participants may use either familiarity-based or proactive strategies; Braver, Gray, & Burgess, 2007; Kane, Conway, Miura, & Colflesh, 2007).

A second and related point is that combining standard subtractive contrasts of different experimental conditions with correlational analyses of individual differences can help distinguish neural processes that are causally involved in modulating overt behavior from those that are epiphenomenal with respect to overt behavior and may even *result* from behavioral changes. A common strategy in fMRI studies is to demonstrate that an experimental manipulation elicits changes in both overt behavior and brain activation, and to then causally attribute the former to the latter. For example, the view that the left inferior frontal gyrus plays a role in resisting cognitive interference during goal-directed processing is based largely on observations that activation in left IFG increases in proportion to the amount of proactive interference present in the environment (for review, see Jonides & Nee, 2006). Such evidence cannot conclusively rule out the possibility that left IFG activation tracks some epiphenomenal cognitive process that covaries with the amount of interference present on a trial but does not itself influence the behavioral response (e.g., familiarity of the interfering stimulus). The causal inference is more strongly supported if it can be shown that changes in IFG activation correlate in meaningful ways with individual differences in resistance to proactive interference (e.g., Gray et al., 2003; Nee et al., 2007).

Of course, brain–behavior relationships can be investigated not only at the level of individual differences, but also within-subject, using trial-by-trial differences in behavior to predict corresponding changes in brain activation. For example, if left IFG is involved in resolving proactive interference, it should presumably show greater activation on trials when participants successfully resist interference than on trials on which they succumb to interference. Integrative approaches that model brain–behavior relationships both within and across subjects should be encouraged, because they can provide even more sophisticated inferences. For example, significant correlations with performance across subjects but not within subjects may indicate that individuals are using qualitatively different strategies that are relatively stable within-session. In a recent fMRI study of decision-making under uncertainty in which participants repeatedly chose between high-probability small rewards and low-probability large rewards (e.g., 90% of 100 points vs. 25% chance of 400 points), we found that activation in PPC correlated strongly with "rational" choice (i.e., selection of the higher expected value) across individuals but not on a trial-by-trial basis (Yarkoni & Braver, 2008). This finding likely reflects the fact that participants who explicitly computed the expected value of each reward – an ability that depends on mental arithmetic operations supported by PPC (Dehaene, Piazza, Pinel, & Cohen, 2005) – were likely to make very few suboptimal choices, whereas those who used heuristic strategies (e.g., selecting the higher-probability reward) did not compute expected value and hence did not activate PPC. The interpretation would have been very different had the PPC covaried with rational choice within subjects but *not* across subjects. In the latter case, a more plausible interpretation would be that participants tended to all use similar computational strategies, and that decision-making performance was largely a function of trial-specific variables (e.g., the difficulty of the trial, random fluctuations in internal noise levels, etc.).

In sum, the relationship between within-subject and individual differences analyses is complex, and the two kinds of analyses should not be viewed simply as two sides of the same coin. Individual

differences analyses offer researchers more than just a second shot at detecting a hypothesized effect. An integrative approach that complements standard experimental contrasts with individual differences analyses as well as trial-by-trial correlations with behavior can potentially provide researchers with sophisticated and powerful insights into brain–behavior relationships that would be difficult to achieve by other means.

Methodological and Statistical Considerations

Power and Sample Size

A paramount consideration when conducting virtually any kind of psychological experiment is statistical power, or the probability that a statistical test will produce a significant result in cases where the tested effect is indeed present in the population. Power is a direct function of three parameters: effect size, statistical threshold, and the sensitivity of the data, which in turn depends on measurement reliability and sample size. Measurement reliability is discussed in greater detail in the next section; for present purposes, we focus on the role of sample size as it is the parameter most easily manipulated by investigators, and arguably the one that presents neuroimaging investigators with the greatest problems.

fMRI studies of individual differences are consistently underpowered. The importance of ensuring that a study has adequate statistical power should be readily apparent. If power is simply the ability to detect the effect one is looking for, why bother conducting a study that has little chance of producing significant results? Yet failures to consider power prior to conducting experimental investigations are surprisingly common in the psychological sciences, as a result of which power levels are often dismally low (for recent discussion, see Maxwell, 2004).

To date, relatively few studies have explored issues related to statistical power in functional neuroimaging studies. The majority of such studies have focused on power-related issues in the context of fMRI design choices – e.g., the trade-off between the ability to *detect* a specific hemodynamic response function and the ability to successfully *estimate* the shape of that response (e.g., Friston, Holmes, Poline, Price, & Frith, 1996; Liu, Frank, Wong, & Buxton, 2001). Only a handful of studies have attempted to determine the number of subjects required in order to ensure an adequate level of power in mixed-model fMRI analyses (Desmond & Glover, 2002; Mumford & Nichols, 2008; Murphy & Garavan, 2004; Thirion et al., 2007), and these studies have focused exclusively on within-subject analyses (e.g., paired *t*-tests contrasting activation in two experimental conditions). No study we know of has attempted to generate power estimates for individual differences analyses in fMRI studies. This absence is problematic, because individual differences analyses have lower power to detect effects than within-subject analyses (assuming that effect size, statistical threshold, and sample size are held constant), and one cannot simply generalize recommendations made specifically for within-subject analyses to the domain of individual differences.

To provide estimates of the sample sizes needed to ensure adequate statistical power in individual differences fMRI analyses, we generated power curves for a standard correlation test (Pearson's *r*) using parameters that are realistic for fMRI studies (Fig. 6.1). For comparison purposes, equivalent curves are presented for a paired *t*-test analysis. Table 6.1 provides point estimates of power level for these two statistical tests at several different effect sizes, sample sizes, and statistical significance thresholds. The figure and table support two main conclusions. First, correlational analyses have substantially lower power to detect effects than one-sample *t*-tests under realistic circumstances. For example, a one-sample *t*-test has 92% power to detect a "large" effect size of $d = 0.8$ in a sample

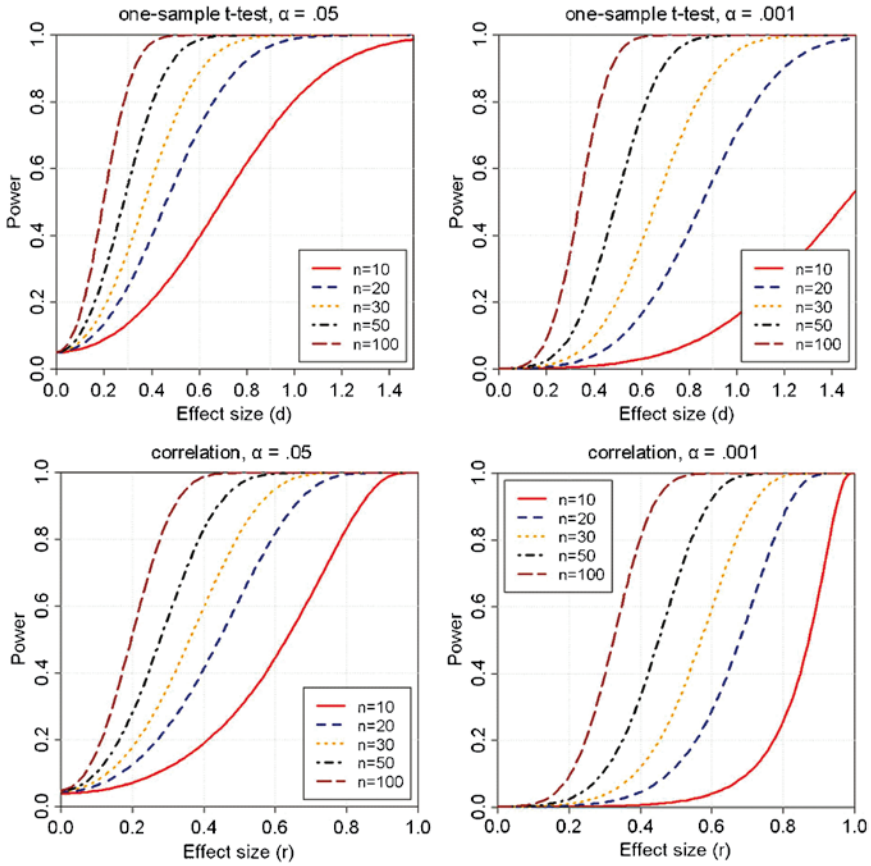

Fig. 6.1 Statistical power as a function of test type (*top* = one-sample *t*-test, *bottom* = Pearson's *r*), sample size, and effect size

of 20 subjects at $p < 0.05$. Yet a d of 0.8 is equivalent to an r of approximately 0.37,[2] and power to detect a correlation of this size at $p < 0.05$ in the same sample is only 35%. Achieving a conventionally adequate power level of 0.8 for this correlational test would require a sample size of 55 – more than double that recommended by power studies that have focused on within-subject analyses (typically about $n = 25$; Desmond & Glover, 2002; Murphy & Garavan, 2004; Thirion et al., 2007).

These analyses make clear that the majority of fMRI analyses of individual differences have very little power to detect all but the most powerful correlational effects. A cursory review of fMRI studies of individual differences in WM and executive control suggests that sample sizes of approximately 15–20 are the norm in previous research. While some studies have used much larger samples (Gray et al., 2003; Hester et al., 2004; Schaefer et al., 2006; Yarkoni, Gray, et al., submitted), others have reported results based on samples as small as 6–9 individuals (Callicott et al., 1999; Fiebach, Rissman, & D'Esposito, 2006; Mattay et al., 2000; Rypma & D'Esposito, 1999). Moreover, many studies that employed individual differences analyses have relied at least in part on whole-brain analyses, necessitating conservative statistical thresholds on the order of $p < 0.001$ or less. When one considers that a correlation test has only 12% power to detect even a "large" correlation of 0.5 at

[2]Cohen's (1988, p. 23) formula provides $r \approx d / \sqrt{d^2 + 4}$.

Table 6.1 Statistical power for correlation and one-sample *t*-test

	Effect size measure		Correlation	One-sample	Relative
n	*r*	*d*	power	*t*-test power	decrease (%)
$\alpha = 0.05$					
10	0.1	0.2	0.05	0.09	44.4
	0.3	0.63	0.12	0.43	72.1
	0.5	1.15	0.3	0.9	66.7
	0.7	1.96	0.63	1	37.0
20	0.1	0.2	0.06	0.13	53.8
	0.3	0.63	0.24	0.76	68.4
	0.5	1.15	0.62	1	38.0
	0.7	1.96	0.95	1	5.0
30	0.1	0.2	0.08	0.19	57.9
	0.3	0.63	0.36	0.91	60.4
	0.5	1.15	0.82	1	18.0
	0.7	1.96	0.99	1	1.0
50	0.1	0.2	0.1	0.28	64.3
	0.3	0.63	0.56	0.99	43.4
	0.5	1.15	0.97	1	3.0
	0.7	1.96	1	1	0.0
$\alpha = 0.001$					
10	0.1	0.2	0	0	–
	0.3	0.63	0	0.03	100.0
	0.5	1.15	0.02	0.25	92.0
	0.7	1.96	0.1	0.85	88.2
20	0.1	0.2	0	0.01	100.0
	0.3	0.63	0.01	0.2	95.0
	0.5	1.15	0.12	0.87	86.2
	0.7	1.96	0.57	1	43.0
30	0.1	0.2	0	0.01	100.0
	0.3	0.63	0.04	0.44	90.9
	0.5	1.15	0.3	0.99	69.7
	0.7	1.96	0.88	1	12.0
50	0.1	0.2	0	0.03	100.0
	0.3	0.63	0.11	0.82	86.6
	0.5	1.15	0.67	1	33.0
	0.7	1.96	1	1	0.0

$p < 0.001$ in a sample size of $n = 20$, it becomes clear that the typical fMRI study of individual differences in WM has little hope of detecting many, if not most, meaningful effects.[3] Thus, there are strong incentives for researchers to use larger samples and theoretically driven analyses that allow ROI-level hypotheses to be tested at more liberal thresholds.

Small samples and effect size inflation. An underappreciated but important consequence of using small sample sizes is that the effect size of significant results may be grossly inflated (Bradley, Smith, & Stoica, 2002; Muncer, Craigie, & Holmes, 2003). When sample sizes are small enough and/or population effects weak enough, the critical value for a statistical test may be much higher than the actual population effect size, so the only way to detect a significant effect is to capitalize

[3]Note that values of $r \geq 0.5$ are extremely rare in most areas of psychology; see Meyer et al. (2001) for a review indicating that most effects across broad domains of psychology and medicine are in the range of 0.1–0.3.

on chance. Failure to consider the possibility of such inflation can then have a number of deleterious effects. First, researchers may grossly overestimate the importance of their findings. Second, the combination of inflated effect sizes and low statistical power may lead researchers to erroneously conclude not only that observed effects are very strong but also that they are highly *selective*. Suppose for example that performance on a WM task shows a uniform correlation of 0.3 with activation across the entire brain in the general population. In a sample of 20 subjects, power to detect a significant correlation for any given observation is only 1.4% at a whole-brain α-level of 0.001. Thus, in a brain comprised of 100,000 voxels, a whole-brain analysis will detect only ~1,400 significant voxels on average – and the mean significant *r* value in these voxels will be highly inflated, at approximately 0.73.[4] An uncritical researcher examining the threshold output from a statistical test could then easily conclude that task performance is very strongly correlated with activation in highly selective brain regions, when in fact the actual population effect is much weaker and shows no spatial selectivity whatsoever.

Finally, inflated effect sizes may lead to misallocation of considerable resources as replications and extensions of a reported "large" effect are unsuccessfully attempted. As studies accumulate, the literature addressing a given research question may come to appear full of "mixed results," with some studies detecting an effect and others failing to, or detecting effects of very different sizes. If one allows that initial reports of correlational effects sometimes grossly overestimate the size of the effect, it should not be surprising if subsequent studies with similar or even larger sample sizes often fail to replicate the original effect, or successfully replicate the effect with a much smaller effect size. Consider, for example, two recent studies by Gray and colleagues (Gray & Braver, 2002; Gray et al., 2005). In a first study, the authors identified strong correlations (0.63–0.84) between participants' BAS scores (measures of behavioral approach tendencies, respectively) and caudal ACC activation in several WM conditions (Gray & Braver, 2002). Yet, in a follow-up study (Gray et al., 2005), with a much larger sample size ($N > 50$), the same correlations were at a much lower level (0.28–0.37). Had a smaller sample size been used in the follow-up study, even a partial replication would have been exceedingly unlikely, and the authors might then have concluded that the initial set of findings were false positives, or reflected some unknown experimental difference. Yet, the seeming discrepancy is easily explained by noting that effect sizes in the first experiment were likely inflated due to small sample size ($n = 14$). In fact, the correlation achieved in the follow-up study was well within the confidence intervals of the effect in the original study but led to a qualitatively different assessment regarding the importance of the personality effect.

In sum, a realistic assessment of power in fMRI studies of individual differences can help researchers avoid undue excitement about the strength and selectivity of observed effects, can provide a realistic assessment of the likelihood of replicating previous effects, and may help explain away many instances of seemingly conflicting findings.

Data peeking and Type I error inflation. A final issue related to the use of small samples in fMRI studies concerns the tendency of fMRI researchers to "peek" at the data – that is, to periodically reanalyze the data every time one or more subjects are added to the sample. Because fMRI data collection and processing is laborious and extremely expensive, the practice of data peeking may seem to make good practical sense. If a targeted effect is already present in a small sample, why bother increasing the sample size at great expense?

While the act of inspecting the data is not harmful in and of itself, the decision to *terminate* data collection as soon as significant results are obtained is rarely if ever defensible. Early termination can lead to substantial inflation of Type I error, because a researcher may erroneously accept a spurious effect that would have gone away if data collection had continued (for discussion of this issue, see Armitage, McPherson, & Rowe, 1969; Pocock, 2006; Strube, 2006; Wagenmakers, 2007).

[4]For the sake of simplicity, this example assumes that each voxel represents an independent observation.

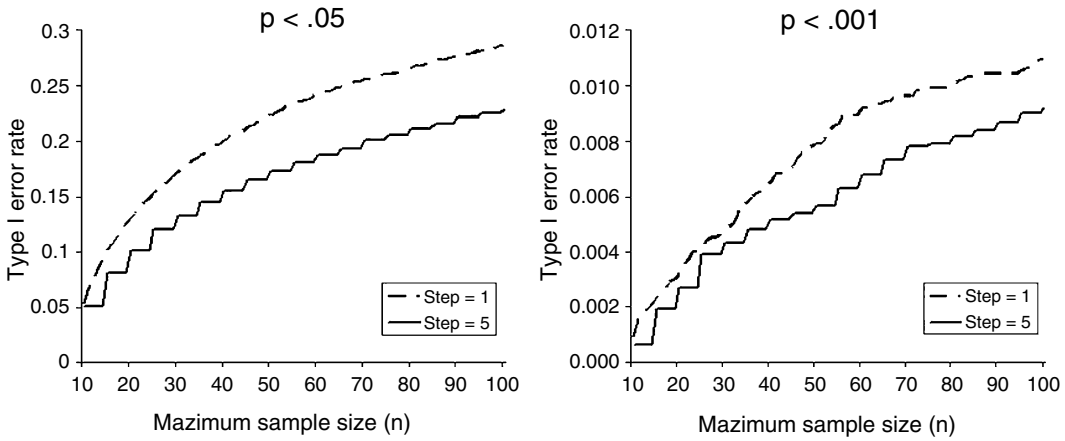

Fig. 6.2 Effects of data peeking on Type I error rate as a function of statistical threshold, sample size, and peeking interval (1 vs. 5). Peeking is assumed to begin after ten participants; beginning earlier would further inflate false positives

Figure 6.2 plots the Type I error rate associated with data peeking for common sample sizes and statistical thresholds (cf. Strube, 2006). Inflation of Type I error is considerable in virtually all cases. For example, given a maximum sample size of 30 and an α-level of 0.05, the probability of falsely detecting a significant effect is over 17% if peeking begins after ten subjects, the correlation is computed every time a subject is added to the sample, and one assumes that the study terminates as soon as a significant effect is found. At a conventional whole-brain threshold of $p<0.001$, the corresponding Type I error rate for the same sample is approximately 0.5% – still low in absolute terms, but a fivefold increase over the nominal rate. If data peeking occurs at less frequent intervals, the inflation factor decreases; however, it remains unacceptably high even at a peeking interval of five subjects (Type I error rate = ~13.4% at $p<0.05$ and ~0.4% at $p<0.001$).

Practical recommendations. Practically speaking, what can researchers do to address the above power-related concerns? We make six recommendations:

1. *Use larger samples*. Whenever possible, the simplest and best cure for a lack of power is to increase the sample size. Use of small samples is often justified anecdotally by noting that effect sizes in fMRI research are very large, making their detection relatively easy. However, as the power curves and Monte Carlo simulations presented above show, there is reason to doubt such assertions. The fact that small-sample studies routinely obtain correlations of 0.7–0.9 almost certainly reflects massive effect size inflation rather than unusually large brain–behavior correlations in the population. Simply put, researchers who are serious about investigating individual differences with fMRI must be willing to collect larger samples. Based on the results presented here, we would suggest $n=40$ as a reasonable lower limit for a study focused *primarily* on individual differences analyses. This sample size is still suboptimal from a power standpoint but may represent an acceptable trade-off between the need to ensure adequate power while limiting data collection costs. However, this generalization should be taken with a grain of salt, and is no substitute for study-specific power calculations.
2. *Perform (and report) power calculations*. There is little reason not to conduct power calculations prior to beginning fMRI data collection. Many power analysis tools are freely available either as stand-alone applications or as add-ons to popular statistics packages,[5] and several websites

[5] We used *R* and the add-on *pwr* library to perform all of the calculations and simulations reported in this chapter.

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 are enormous. There is no good reason we can see for failing to report power calculations whenever individual differences analyses are reported in fMRI studies. Reviewers and editors should similarly be encouraged to request or require authors to report power calculations when none are provided.

3. *Test* a priori *hypotheses whenever possible*. While whole-brain analyses are an important complement to any set of focused regional analyses, the use of ROI-level tests allows correlations to be tested at much more liberal statistical thresholds (e.g., $p < 0.05$ instead of $p < 0.001$). Of course, testing too many hypotheses at a regional level can introduce its own multiple-comparisons problem, and the privileged status of theoretical hypotheses should not be abused by testing several dozen regions without further controlling the Type I error rate.
4. *Do not base sample size decisions on inspection of provisional data.* While some amount of data peeking for quality assurance purposes may be inevitable in fMRI studies, researchers should peek only to ensure to adequacy of the fMRI signal itself, and not the adequacy of the results in relation to the hypothesis. Researchers should not cease collecting fMRI data once sufficiently interesting results are obtained, nor should they chase a marginal correlation with additional subjects until it becomes significant. The recommended approach is to select a sample size based on a priori power calculations and stick to it. Deviating from the chosen sample size based on preliminary results will result in considerable inflation of Type I error.
5. *Avoid attributing null results or replication failures to experimental factors without good cause*. It is often tempting to attribute a replication failure to some minute difference between the original and present studies. However, unless sample size is very large and/or the expected effects are extremely strong, by far the most plausible explanations for a null result in an individual differences fMRI analysis are that (a) the present null result represents a Type II error or (b) the original result represented a Type I error.
6. *Pay little attention to the apparent strength or spatial selectivity of correlational effects.* Small sample sizes may dramatically inflate the size of significant effects, and low power tends to induce an illusion of spatial selectivity. Consequently, a good deal of skepticism should be applied when interpreting small-sample size correlational results that appear to be very strong and highly selective.

Reliability

Reliability refers to the ability of an instrument or measure to produce consistent results when tested under similar conditions. Measurement reliability is a paramount concern in virtually all empirical investigations because the reliability of a measure sets an upper bound on the extent to which its variance is available to correlate with other measures (Cohen, West, Aiken, & Cohen, 2002). The detectable correlation between the *observed* scores of two variables is formally equivalent to the square root of the product of the two reliabilities multiplied by the true population correlation. Attenuation of effects due to measurement unreliability may have few practical consequences in cases where effects and/or sample sizes are very large because, so long as an effect is detectable and estimates of measurement reliability are available, it is always possible to correct for unreliability and estimate the true population effect size. However, as the above discussion of power should make clear, fMRI studies rarely occupy this privileged niche. Given the small sample sizes typical of fMRI studies, a difference between population-level and sample-level effect sizes could easily amount to the difference between a significant effect and a null result. Thus, the viability of individual differences analyses in fMRI studies depends critically on ensuring that the measures used – whether of behavior or brain function – are sufficiently reliable.

Behavioral measures. By convention, reliability coefficients of 0.8 are considered adequate in most domains of psychological research, and coefficients of 0.9 or more are generally considered high. Given that most fMRI analyses of individual differences in WM and executive control use behavioral measures (e.g., task performance or cognitive ability measures) to predict brain activity, one might expect a similar convention to hold in the fMRI literature. Unfortunately, reliability coefficients for behavioral measures are rarely reported in fMRI studies. This absence is of little concern in studies that use standard measures that have been psychometrically validated – e.g., the RAPM or Engle and colleagues' version of the Operation Span task (Unsworth, Heitz, Schrock, & Engle, 2005) – because it is unlikely that the reliability of such measures will depart radically from published norms for any given study. However, many fMRI studies use nonstandard individual differences measures as predictors of brain activation, e.g., measures of RT or response accuracy derived from idiosyncratically implemented or entirely novel cognitive tasks. In the latter case, it is important to supplement any reported correlational results with an estimate of the reliability of the measures involved. Given that reliability can be easily estimated in a variety of ways (e.g., virtually all major statistical packages can easily compute Cronbach's α, the most common measure of internal consistency), researchers should be encouraged to report a coefficient of reliability for all individual differences measures used to predict brain activity – a step only a few fMRI studies have taken (e.g., Gillath, Bunge, Shaver, Wendelken, & Mikulincer, 2005; Heinz et al., 2004).

Neuroimaging measures. In contrast to behavioral measures, it is not quite as straightforward for researchers to estimate the reliability of the BOLD signal. Subjects are typically tested on only one occasion in most studies, and measurements at different voxels or scanning frames cannot be thought of as items of a measure in the same sense as questions on a questionnaire. The most viable route to determining reliability given such constraints is to compute some form of split-half reliability coefficient, e.g., by comparing activation across different runs within the same session, or by randomly coding events of a single type using two different variables. While this approach is admittedly time-consuming, and produces results that may be difficult to interpret and report because reliability coefficients may vary considerably from voxel to voxel and experimental condition to experimental condition, we believe the potential benefits are sufficiently great to warrant wider adoption.

General estimates of the reliability of fMRI data can be gleaned from a number of studies that have explicitly sought to quantify the reliability of the BOLD signal.[6] Overall, such studies provide a mixed picture. On the positive side, several studies that assessed test–retest reliability at varying intervals reported adequate or even high reliability coefficients across a range of experimental tasks (e.g., Aron et al., 2006; Fernandez et al., 2003; Specht et al., 2003). For example, Aron et al. (2006) scanned eight participants who performed a classification learning task on two separate occasions 1 year apart and found intraclass correlation coefficients (ICCs) greater than 0.8 in many voxels that were activated at the group level. Such results clearly demonstrate that it is *possible* to measure brain activation reliably with fMRI; however, there is no guarantee that this conclusion will generalize across different samples, designs, scanners, and experimental contrasts.

[6]Note that the term *reliability* is used here to refer specifically to the stability of the rank order of BOLD activation across subjects. That is, the BOLD signal can be considered reliable if individuals who show high levels of activation when scanned on one occasion also show high levels of activation when scanned on another occasion under the same conditions. The term reliability is also often used to refer to the *replicability* or *reproducibility* of fMRI results at the group level – e.g., deeming the BOLD signal reliable if approximately the same pattern of group-level activation can be replicated across different samples, scanners, institutions, task variants, etc. Although these two senses of reliability are interrelated, they are not equivalent. We focus here only on the former sense, as it is the one relevant for individual differences analyses.

Indeed, other studies have provided decidedly less optimistic results (Johnstone et al., 2005; Manoach et al., 2001; Manuck, Brown, Forbes, & Hariri, 2007). For example, Johnstone et al. 2005 investigated the test–retest reliability of amygdala reactivity to neutral and fearful faces across three testing occasions spanning 8 weeks. ICCs peaked around 0.8 in the left amygdala, but were near zero in the right amygdala for several contrasts. Manuck et al. 2007 conducted a similar study but with a longer (1 year) test–retest interval. In contrast to Johnstone et al.'s results, Manuck et al. found moderate stability in the right amygdala ($r=0.59$) but no stability in the left amygdala ($r=-0.08$). Note that low test–retest reliability does not necessarily imply low reliability in general as test–retest reliability estimates can differ considerably from estimates of internal consistency. That is, it is possible for regional brain activation to show adequate reliability on one occasion yet fail to show consistency over time due to the systematic influence of other factors (e.g., participants' mood). Nonetheless, the complete absence of temporal consistency in some analyses is cause for concern and raises the worrisome possibility that correlational effects may be impossible to detect in some regions and experiments.[7] Moreover, even in studies that have identified highly reliable activations, the locus of reliable signal tends to be circumscribed to those regions that show significant group-level involvement in the task (Aron et al., 2006; Specht et al., 2003), potentially limiting the scope of individual differences analyses even further.

Increasing reliability. Considerable work remains to be done in order to better understand the conditions that affect the reliability of individual differences in the BOLD signal. In the interim, several steps can be taken in the hopes of increasing reliability. First, researchers should take care to use reliable behavioral measures as predictors of brain activity and should report reliability coefficients for such measures whenever possible. Second, in cases where the process is not too laborious, researchers can obtain rough estimates of BOLD signal reliability by conducting split-half analyses as noted above. Third, reliability can be increased by employing data reduction or latent variable techniques such as principal components analysis (PCA) or SEM that "triangulate" on reliable variance. For example, factor analytic techniques can be used to extract a small number of latent activation variables from a large number of ROIs or behavioral variables (Badre, Poldrack, Paré-Blagoev, Insler, & Wagner, 2005; Peterson et al., 1999; Wager, Sylvester, et al., 2005; Yarkoni, Gray, et al., submitted) or from a single ROI across different contrasts (Yarkoni, Gray, et al., submitted). Such factors, which reflect only common (and therefore, reliable) variance, are more likely to correlate with behavioral variables, other things being equal. (A secondary benefit of such an approach is the reduction in the number of statistical comparisons tested.)

Finally, researchers who are specifically interested in individual differences and are willing to sacrifice some power to detect within-subject effects may be able to increase the reliability of brain activation measures by deliberately selecting experimental tasks that produce highly variable performance across individuals. For example, if one's goal is to investigate the neural correlates of individual differences in WM capacity, maximal reliability is likely to be attained when the experimental task used in the scanner is relatively difficult, and a wide range of performance levels is observed. Conversely, when a task is relatively easy and performance is near ceiling levels for most participants, as is common in many fMRI studies of WM, activation in regions associated with task performance may be relatively unreliable because the amount of effort a participant invests may have little influence on their performance level (i.e., one could invest a little effort or a lot, and the behavioral data would not reflect this variability).

[7]It may also be that these low test-retest correlations reflect a genuine lack of stable individual differences. However, the strong hemispheric asymmetry and conflicting findings across studies seem to weigh against such a conclusion, as does the fact that numerous studies have detected individual differences effects in the amygdala using similar contrasts (Canli, Sivers, Whitfield, Gotlib, & Gabrieli, 2002; Canli et al., 2001).

Outliers

One of the most important topics in statistical research methodology concerns how researchers should identify and deal with outliers – extreme observations that fall outside the expected bounds of a distribution. In addition to standard reasons for worrying about outliers – that they skew the distribution, inflate the variance, and can bias results – there are at least three additional reasons why outliers may present particular cause for concern in the domain of neuroimaging. First, neuroimaging sample sizes are heavily constrained by the high cost of data collection. Thus, the most general strategy for reducing outlier influence – increasing sample size – is often not viable. Second, neuroimaging datasets are highly susceptible to various forms of artifact, potentially resulting in a disproportionate number of outliers (Ojemann et al., 1997; Wager, Keller, Lacey, & Jonides, 2005). Third, the large number of statistical comparisons performed in typical neuroimaging analyses can easily result in a failure to detect outliers, since it is not possible to visually inspect a scatter plot for each comparison of interest as behavioral researchers commonly do. When reporting significant correlations in specific brain regions, researchers often display the corresponding scatter plots and explicitly discuss the potential role of outliers if they are present. This approach works well for voxels or regions that a statistical test or prior hypothesis have indicated are worth examining, but it is of no help in cases where a real correlation exists but is obscured by an outlier that biases the regression coefficient *away* from significance and consequently *fails* to be detected. The need to preemptively identify and control for the influence of outliers therefore calls for the use of procedures other than the standard parametric tests.

Identifying outliers. Numerous methods exist for visually or quantitatively identifying outliers (for review, see Barnett & Lewis, 1994; Iglewicz & Hoaglin, 1993). In practice, the most common approaches are to either label observations as outliers based on a semiarbitrary prespecified criterion (e.g., falling more than N standard deviations from the mean) or to employ the boxplot method, which identifies as outliers any observations that fall a certain distance outside the interquartile range (McGill, Tukey, & Larsen, 1978; Tukey, 1977). These methods are easy to apply but are relatively unprincipled and do not provide guidance concerning how to deal with outliers once they have been identified. A common approach in many areas of psychology (e.g., analyses of RT in cognitive psychology; for review, see Ratcliff, 1993) is to simply remove all observations labeled as outliers from analysis. However, this approach may be inadvisable for individual differences analyses in neuroimaging studies where sample sizes are small and power is already low to begin with; moreover, dropping outliers arbitrarily is liable to bias resulting in regression coefficients (often in ways that, intentionally or not, favor the hypothesis). An alternative is to apply a mathematical transformation to the data (e.g., taking the natural logarithm) so as to alter the shape of the distribution and minimize outlier influence. A disadvantage of the latter approach is that it complicates interpretation of results, since the conceptualization of the transformed variable is often unclear. Whereas a value given in percent change in the BOLD signal is readily interpretable, a log-transformed BOLD change value may not be. Moreover, there is rarely a principled reason to apply a particular mathematical transformation to the data, potentially resulting in a series of post-hoc transformations that can lead to capitalization on chance if researchers are not careful.

Accommodating outliers. A potentially preferable alternative to manipulating the data itself is the use of statistical procedures that can accommodate outliers by reducing their influence. One class of such techniques includes nonparametric tests which make no assumptions about the distributions of the variables being tested. For example, Pearson's correlation coefficient can be replaced with a number of alternatives such as Spearman's rho, Kendall's tau, or bootstrapped correlation coefficients (for reviews, see Chen & Popovich, 2002; Efron & Tibshirani, 1993; Nichols & Holmes, 2002; Siegel & Castellan, 1988). These tests are appropriate in cases where assumptions of normality are violated (as they may be in the presence of outliers); however, they trade off decreased susceptibility to outliers and distribution violations against lower power to detect correlations when the assumption

of normality is met (Siegel & Castellan, 1988). Of course, given the central limit theorem, such normality assumptions are going to be more likely to be violated when the sample size is small ($N<30$), so this issue may be especially important for neuroimaging studies. Nonparametric statistical approaches have been discussed previously in the neuroimaging literature, in terms of these tradeoffs (for a review and discussion of available software tools, see Nichols & Holmes, 2002).

An alternative class of methods, termed *robust* estimation methods, reduce outlier influence by downweighting extreme scores in various ways (Cleveland, 1979; Rousseeuw & Leroy, 1987; Wager, Keller, et al., 2005). In contrast to most nonparametric methods, simulation studies suggest that many robust estimation methods retain nearly the same level of power as parametric methods under conditions of normality while providing the expected increase in power in the presence of outliers. Wager and colleagues recently used both simulated and empirical data to demonstrate that robust estimation techniques can provide substantial power increases in neuroimaging analyses affected by outliers without affecting the false positive rate (Wager, Keller, et al., 2005). Importantly, they show that there is virtually no penalty associated with the use of robust estimation when the data contain *no* outliers. Given appropriate software implementation, widespread use of robust estimation in neuroimaging analysis could therefore provide a relatively principled approach to minimizing the influence of outliers without requiring any special attention or expertise on the part of researchers. The primary limitation of such methods is that they are somewhat more computationally intensive than ordinary least-squares regression.

Conclusions

Individual differences approaches have found their way into the toolkits of an ever-increasing number of cognitive neuroscientists studying WM/EC. Such approaches have the potential to greatly further understanding of WM/EC; however, their use should ideally be guided by an appreciation of their limitations. A number of specific methodological issues should be carefully considered when planning, conducting, or analyzing a study involving individual differences analyses. This chapter considered a number of such issues, including statistical power limitations, effect size inflation, measurement reliability, and treatment of outliers. An overarching theme is that these issues overlap only partially with those that apply to standard within-subject analyses based on experimental contrasts. Likewise, we feel it is important that researchers treat individual differences and within-subject analysis approaches as complementary tools that are differentially sensitive to specific kinds of mechanisms, and not simply as two different ways to test for convergent effects. Although we focused primarily on fMRI studies in the present chapter, the issues we discuss are widely applicable to other cognitive neuroscience methods such as PET, TMS, and EEG/ERP. Our hope is that other researchers will find these considerations useful and take them into account in future when designing studies involving analyses of individual differences, and interpreting their results.

References

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A, 132*, 235–244.

Aron, A. R., Gluck, M. A., & Poldrack, R. A. (2006). Long-term test–retest reliability of functional MRI in a classification learning task. *NeuroImage, 29*, 1000–1006.

Badre, D., Poldrack, R. A., Paré-Blagoev, E. J., Insler, R. Z., & Wagner, A. D. (2005). Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. *Neuron, 47*(6), 907–918.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. Chichester: Wiley.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
Bradley, M. T., Smith, D., & Stoica, G. (2002). A Monte-Carlo estimation of effect size distortion due to significance testing. *Perceptual and Motor Skills, 95*(3), 837–842.
Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *NeuroImage, 5*(1), 49–62.
Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory*. Oxford: Oxford University Press.
Buchsbaum, B. R., Greer, S., Chang, W. L., & Berman, K. F. (2005). Meta-analysis of neuroimaging studies of the Wisconsin card-sorting task and component processes. *Human Brain Mapping, 25*(1), 35–45.
Bunge, S. A., Ochsner, K. N., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (2001). Prefrontal regions involved in keeping information in and out of mind. *Brain, 124*(10), 2074.
Burock, M. A., Buckner, R. L., Woldorff, M. G., Rosen, B. R., & Dale, A. M. (1998). Randomized event-related experimental designs allow for extremely rapid presentation rates using functional MRI. *Neuroreport, 9*(16), 3735–3739.
Cabeza, R., Anderson, N. D., Locantore, J. K., & McIntosh, A. R. (2002). Aging gracefully: Compensatory brain activity in high-performing older adults. *NeuroImage, 17*(3), 1394–1402.
Callicott, J. H., Mattay, V. S., Bertolino, A., Finn, K., Coppola, R., Frank, J. A., et al. (1999). Physiological characteristics of capacity constraints in working memory as revealed by functional MRI. *Cerebral Cortex, 9*(1), 20–26.
Canli, T., Sivers, H., Whitfield, S. L., Gotlib, I. H., & Gabrieli, J. D. (2002). Amygdala response to happy faces as a function of extraversion. *Science, 296*(5576), 2191.
Canli, T., Zhao, Z., Desmond, J. E., Kang, E., Gross, J., & Gabrieli, J. D. (2001). An fMRI study of personality influences on brain reactivity to emotional stimuli. *Behavioral Neuroscience, 115*(1), 33–42.
Caplan, J. B., McIntosh, A. R., & De Rosa, E. (2007). Two distinct functional networks for successful resolution of proactive interference. *Cerebral Cortex, 17*(7), 1650.
Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage.
Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*(368), 829–836.
Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*(3), 249.
Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd Ed.), Lawrence Erlbaum, Hillsdale, NJ.
Cohen, J., West, S. G., Aiken, L., & Cohen, P. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163–183.
Dale, A. M., & Buckner, R. L. (1997). Selective averaging of rapidly presented individual trials using fMRI. *Human Brain Mapping, 5*(5), 329–340.
Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2005). Three parietal circuits for number processing. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 433–455). New York: Psychology Press.
Della-Maggiore, V., Sekuler, A. B., Grady, C. L., Bennett, P. J., Sekuler, R., & McIntosh, A. R. (2000). Corticolimbic interactions associated with performance on a short-term memory task are modified by age. *Journal of Neuroscience, 20*(22), 8410.
Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods, 118*(2), 115–128.
Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences, 23*(10), 475–483.
Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.
Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309–331.
Fernandez, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., et al. (2003). Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. *Neurology, 60*(6), 969–975.
Fiebach, C. J., Rissman, J., & D'Esposito, M. (2006). Modulation of inferotemporal cortex activation during verbal working memory maintenance. *Neuron, 51*(2), 251–261.
Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences, 102*(27), 9673–9678.
Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General, 133*(1), 101–135.

Friston, K. J., Holmes, A., Poline, J. B., Price, C. J., & Frith, C. D. (1996). Detecting activations in PET and fMRI: Levels of inference and power. *NeuroImage, 4*(3), 223–235.

Geake, J. G., & Hansen, P. C. (2005). Neural correlates of intelligence as revealed by fMRI of fluid analogies. *NeuroImage, 26*(2), 555–564.

Gibbs, S. E., & D'Esposito, M. (2005). Individual capacity differences predict working memory performance and prefrontal activity following dopamine receptor stimulation. *Cognitive, Affective & Behavioral Neuroscience, 5*(2), 212–221.

Gibbs, S. E. B., & D'Esposito, M. (2006). A functional magnetic resonance imaging study of the effects of pergolide, a dopamine receptor agonist, on component processes of working memory. *Neuroscience, 139*(1), 359–371.

Gillath, O., Bunge, S. A., Shaver, P. R., Wendelken, C., & Mikulincer, M. (2005). Attachment-style differences in the ability to suppress negative thoughts: Exploring the neural correlates. *NeuroImage, 28*, 835–847.

Glabus, M. F., Horwitz, B., Holt, J. L., Kohn, P. D., Gerton, B. K., Callicott, J. H., et al. (2003). Interindividual differences in functional interactions among prefrontal, parietal and parahippocampal regions during working memory. *Cerebral Cortex, 13*(12), 1352–1361.

Grady, C. L., McIntosh, A. R., Bookstein, F., Horwitz, B., Rapoport, S. I., & Haxby, J. V. (1998). Age-related changes in regional cerebral blood flow during working memory for faces. *NeuroImage, 8*(4), 409–425.

Gray, J. R., & Braver, T. S. (2002). Personality predicts working-memory-related activation in the caudal anterior cingulate cortex. *Cognitive, Affective & Behavioral Neuroscience, 2*(1), 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.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6*(3), 316–322.

Haier, R. J., Siegel, B. V., Jr., MacLachlan, A., Soderling, E., Lottenberg, S., & Buchsbaum, M. S. (1992). Regional glucose metabolic changes after learning a complex visuospatial/motor task: A positron emission tomographic study. *Brain Research, 570*(1–2), 134–143.

Haier, R. J., White, N. S., & Alkire, M. T. (2003). Individual differences in general intelligence correlate with brain function during nonreasoning tasks. *Intelligence, 31*(5), 429–441.

Heinz, A., Siessmeier, T., Wrase, J., Hermann, D., Klein, S., Grusser-Sinopoli, S. M., et al. (2004). Correlation between dopamine D2 receptors in the ventral striatum and central processing of alcohol cues and craving. *American Journal of Psychiatry, 161*(10), 1783–1789.

Hester, R., Fassbender, C., & Garavan, H. (2004). Individual differences in error processing: A review and reanalysis of three event-related fMRI studies using the GO/NOGO task. *Cerebral Cortex, 14*(9), 986–994.

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers*. Milwaukee, WI: ASQ Quality Press.

Johnstone, T., Somerville, L. H., Alexander, A. L., Oakes, T. R., Davidson, R. J., Kalin, N. H., et al. (2005). Stability of amygdala BOLD response to fearful faces over multiple scan sessions. *NeuroImage, 25*, 1112–1123.

Jonides, J., & Nee, D. E. (2006). Brain mechanisms of proactive interference in working memory. *Neuroscience, 139*(1), 181–193.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 615–622.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*(2), 189–21728.

Kim, J., Zhu, W., Chang, L., Bentler, P. M., & Ernst, T. (2007). Unified structural equation modeling approach for the analysis of multisubject, multivariate functional MRI data. *Human Brain Mapping, 28*(2), 85–93.

Kimberg, D. Y., Aguirre, G. K., Lease, J., & D'Esposito, M. (2001). Cortical effects of bromocriptine, a D-2 dopamine receptor agonist, in human subjects, revealed by fMRI. *Human Brain Mapping, 12*(4), 246–257.

Kimberg, D. Y., D'Esposito, M., & Farah, M. J. (1997). Effects of bromocriptine on human subjects depend on working memory capacity. *Neuroreport, 8*(16), 3581.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.

Kondo, H., Morishita, M., Osaka, N., Osaka, M., Fukuyama, H., & Shibasaki, H. (2004). Functional roles of the cingulo-frontal network in performance on working memory. *NeuroImage, 21*(1), 2–14.

Kumari, V., Ffytche, D. H., Williams, S. C., & Gray, J. A. (2004). Personality predicts brain responses to cognitive demands. *Journal of Neuroscience, 24*(47), 10636.

Larson, G. E., Haier, R. J., LaCasse, L., & Hazen, K. (1995). Evaluation of a "mental effort" hypothesis for correlations between cortical metabolism and intelligence. *Intelligence, 21*(3), 267–278.

Lee, K. H., Choi, Y. Y., Gray, J. R., Cho, S. H., Chae, J. H., Lee, S., et al. (2006). Neural correlates of superior intelligence: Stronger recruitment of posterior parietal cortex. *NeuroImage, 29*(2), 578–586.

Liu, T. T., Frank, L. R., Wong, E. C., & Buxton, R. B. (2001). Detection power, estimation efficiency, and predictability in event-related fMRI. *NeuroImage, 13*(4), 759–773.

Locke, H. S., & Braver, T. S. (2008). Motivational influences on cognitive control: Behavior, brain activation, and individual differences. *Cognitive, Affective & Behavioral Neuroscience, 8*(1), 99–112.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288*(5472), 1835.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*(1), 83–104.

Manoach, D. S., Halpern, E. F., Kramer, T. S., Chang, Y., Goff, D. C., Rauch, S. L., et al. (2001). Test-retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. *American Journal of Psychiatry, 158*(6), 955–958.

Manuck, S. B., Brown, S. M., Forbes, E. E., & Hariri, A. R. (2007). Temporal stability of individual differences in amygdala reactivity. *American Journal of Psychiatry, 164*(10), 1613.

Mattay, V. S., Callicott, J. H., Bertolino, A., Heaton, I., Frank, J. A., Coppola, R., et al. (2000). Effects of dextroamphetamine on cognitive performance and cortical activation. *NeuroImage, 12*(3), 268–275.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The American Statistician, 32*(1), 12–16.

McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. (1996). Spatial pattern analysis of functional brain images using Partial Least Squares. *NeuroImage, 3*(3), 143–157.

McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances. *NeuroImage, 23*, 250–263.

Mecklinger, A., Weber, K., Gunter, T. C., & Engle, R. W. (2003). Dissociable brain mechanisms for inhibitory control: Effects of interference content and working memory capacity. *Cognitive Brain Research, 18*(1), 26–38.

Mehta, M. A., Owen, A. M., Sahakian, B. J., Mavaddat, N., Pickard, J. D., & Robbins, T. W. (2000). Methylphenidate enhances working memory by modulating discrete frontal and parietal lobe regions in the human brain. *Journal of Neuroscience, 20*(6), RC65.

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*(2), 128–165.

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage, 39*(1), 261–268.

Muncer, S. J., Craigie, M., & Holmes, J. (2003). Meta-analysis and power: Some suggestions for the use of power in research synthesis. *Understanding Statistics, 2*(1), 1–12.

Murphy, K., & Garavan, H. (2004). An empirical investigation into the number of subjects required for an event-related fMRI study. *NeuroImage, 22*(2), 879–885.

Nee, D. E., Jonides, J., & Berman, M. G. (2007). Neural mechanisms of proactive interference-resolution. *NeuroImage, 38*(4), 740–751.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping, 15*(1), 1–25.

Ojemann, J. G., Akbudak, E., Snyder, A. Z., McKinstry, R. C., Raichle, M. E., & Conturo, T. E. (1997). Anatomic localization and quantitative analysis of gradient refocused echo-planar fMRI susceptibility artifacts. *NeuroImage, 6*(3), 156–167.

Omura, K., Aron, A., & Canli, T. (2005). Variance maps as a novel tool for localizing regions of interest in imaging studies of individual differences. *Cognitive, Affective & Behavioral Neuroscience, 5*(2), 252–261.

Osaka, M., Osaka, N., Kondo, H., Morishita, M., Fukuyama, H., Aso, T., et al. (2003). The neural basis of individual differences in working memory capacity: An fMRI study. *NeuroImage, 18*(3), 789–797.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping, 25*(1), 46–59.

Peterson, B. S., Skudlarski, P., Gatenby, J. C., Zhang, H., Anderson, A. W., & Gore, J. C. (1999). An fMRI study of stroop word-color interference: Evidence for cingulate subregions subserving multiple distributed attentional systems. *Biological Psychiatry, 45*(10), 1237–1258.

Pocock, S. J. (2006). Current controversies in data monitoring for clinical trials. *Clinical Trials, 3*(6), 513.

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods, 10*, 178–192.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*(3), 510–532.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and vocabulary scales*. Oxford, UK: Oxford Psychologists Press.

Reuter-Lorenz, P. A. (2002). New visions of the aging mind and brain. *Trends in Cognitive Sciences, 6*(9), 394–400.

Richeson, J. A., Baird, A. A., Gordon, H. L., Heatherton, T. F., Wyland, C. L., Trawalter, S., et al. (2003). An fMRI investigation of the impact of interracial contact on executive function. *Nature Neuroscience, 6*(12), 1323–1328.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rypma, B., Berger, J. S., & D'Esposito, M. (2002). The influence of working-memory demand and subject performance on prefrontal cortical activity. *Journal of Cognitive Neuroscience, 14*(5), 721–731.

Rypma, B., Berger, J. S., Prabhakaran, V., Martin Bly, B., Kimberg, D. Y., Biswal, B. B., et al. (2006). Neural correlates of cognitive efficiency. *NeuroImage, 33*(3), 969–979.

Rypma, B., & D'Esposito, M. (1999). The roles of prefrontal brain regions in components of working memory: Effects of memory load and individual differences. *Proceedings of the National Academy of Sciences, 96*(11), 6558–6563.

Schaefer, A., Braver, T. S., Reynolds, J. R., Burgess, G. C., Yarkoni, T., & Gray, J. R. (2006). Individual differences in amygdala activity predict response speed during working memory. *Journal of Neuroscience, 26*(40), 10120–10128.

Schlösser, R. G. M., Wagner, G., & Sauer, H. (2006). Assessing the working memory network: Studies with functional magnetic resonance imaging and structural equation modeling. *Neuroscience, 139*(1), 91–103.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Specht, K., Willmes, K., Shah, N. J., & Jaencke, L. (2003). Assessment of reliability in functional imaging studies. *Journal of Magnetic Resonance Imaging, 17*(4), 463–471.

Strube, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods, 38*(1), 24–27.

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., & Poline, J. B. (2007). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage, 35*(1), 105–120.

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science, 315*(5811), 515.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498–505.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779–804.

Wager, T. D., Jonides, J., & Reading, S. (2004). Neuroimaging studies of shifting attention: A meta-analysis. *NeuroImage, 22*(4), 1679–1693.

Wager, T. D., Keller, M. C., Lacey, S. C., & Jonides, J. (2005). Increased sensitivity in neuroimaging analyses using robust regression. *NeuroImage, 26*(1), 99–113.

Wager, T. D., Sylvester, C. Y. C., Lacey, S. C., Nee, D. E., Franklin, M., & Jonides, J. (2005). Common and unique components of response inhibition revealed by fMRI. *NeuroImage, 27*(2), 323–340.

Yarkoni, T., Barch, D. M., Gray, J. A., Conturo, T., & Braver, T. S. (2009). BOLD correlates of trial-by-trial reaction time variability in gray and white matter: a multi-study fMRI analysis. *PLoS ONE*, 4, e4527.

Yarkoni, T., & Braver, T. S. (2008). *Dissociable influences of probability, magnitude, and expected value on decision-making.* Paper presented at the Cognitive Neuroscience Society, San Francisco, CA.

Yarkoni, T., Braver, T. S., Gray, J. R., & Green, L. (2005). Prefrontal brain activity predicts temporally extended decision-making behavior. *Journal of the Experimental Analysis of Behavior, 84*(3), 537–554.

Yarkoni, T., Gray, J. R., & Braver, T. S. (submitted). Medial posterior parietal cortex activation predicts working memory performance within and across subjects.