Behavior Research Methods
2009, 41 (4), 1061-1072
doi:10.3758/BRM.41.4.1061

Research design issues for the use of magnetic resonance imaging machines in brain studies of psychological/psychiatric variables

Mark Y. Czarnolewski
Silver Spring, Maryland

The magnetic resonance imaging (MRI) machine itself has an impact on the likelihood of obtaining successful measurements of brain size in certain groups of subjects. The differential selection and attrition in both cross-sectional and longitudinal designs, therefore, indicate that the MRI coincidentally serves as a screen for the anatomical structure of the brains that are successfully scanned. This screening effect introduces confounds in experiments whose very hypotheses are focused on comparing anatomical differences in subjects who differ, for example, in their reactions to anxiety-inducing situations. Here, behavioral interventions and possible statistical models are presented in order to reduce attrition and other effects of the confounds introduced by the MRI measurement process in research. Child and adolescent research—particularly in the attention-deficit/hyperactivity disorder research area—is used as an example to clarify and delineate the general research principles presented in the present article.

A recent article (Eshed, Althoff, Hamm, & Hermann, 2007) in the *Journal of Magnetic Resonance Imaging* documented both the impact of the MRI environment and the relationship of demographic variables and anatomical site variables on the likelihood of obtaining successful MRI scans. First, the present article builds on prior efforts (Czarnolewski, 2005) and reports a sampling of the literature documenting the impact the MRI environment has on subjects' being able to tolerate the scanning process, which, in turn, affects whether successfully completed and unblurred scans are obtained. Second, the present article will detail the research design issues affecting the study of psychological and psychiatric variables when researchers employ MRIs, including a discussion of the corresponding threats to internal and external validity. The article also presents such threats to accepted standards in the medical and epidemiological fields, standards that suggest data reporting requirements and experimental strategies in studies experiencing attrition like that experienced with MRIs. Third, making use of the pertinent literature, the article will present data analysis strategies in conjunction with behavioral intervention techniques for addressing attrition. The relevance of these issues to child and adolescent research—most notably, to the attention-deficit/hyperactivity disorder (ADHD) research area—will be highlighted. Admittedly, much of the presentation in this article applies to other technologies, such as functional magnetic resonance imaging (fMRI) in MRIs (Pearlson & Calhoun, 2007); however, much of the literature documenting subject reactions to primarily closed MRIs focuses on structural MRIs. Consequently, this will be the focus of the section on subject reactions.

Some Reactions to MRIs, As Described in the Literature

The structural MRI machine has become a widely used tool for comparing brain region sizes between diagnostic and control groups. However, because of differential subject reactions to the MRI experience (Eshed et al., 2007; Fitz, 1989; Lukins, Davan, & Drummond, 1997; Meléndez & McCrank, 1993; Murphy & Brunberg, 1997; Rosenberg et al., 1997), the present article contends that this adverse experience coincidentally serves as a screen for the anatomical structure of successfully scanned brains. This section will present results from clinical populations and research studies.

Eshed et al. (2007) found that women were more likely to stop their MRI scans than men were, and that an MRI scan of the head was more likely to result in premature demands for cessation by the patient. Rosenberg et al. (1997) stated that the experiences of MRI physicians, technologists, and nursing staff suggest that pediatric subjects require more preparation for MRI exams than do adults. They cited Fitz (1989), who reported the following rates of children requiring sedation for older, closed clinical MRI procedures: 50% of 6-year-olds, more than 30% of 7- and 8-year-olds, and 10% of 9- to 12-year-olds. Rosenberg et al. hypothesized that higher rates of necessary sedation might be expected with pediatric psychiatric patients—especially with pediatric patients suffering from anxiety.

A brief presentation of some of the literature reporting reactions to MRIs may help one to better identify the subject variables corresponding to higher attrition rates in general, as well as in particular diagnostic groups, such as children with ADHD. A particular concern is that

M. Y. Czarnolewski, mczarnol@yahoo.com

variables (such as claustrophobia) may be more common in diagnostic groups than in control groups. In general, approximately 30% of patients experience anxiety reactions, and 5% to 10% of patients experience severe panic or claustrophobia (Meléndez & McCrank, 1993). There is a positive correlation between anxiety experienced during an MRI scan and the development of MRI-related fears when patients take follow-up scans 7 or more months later (Lukins et al., 1997). These authors also found an increase in fear of MRIs, even in the patients who were assigned to groups undergoing relaxation interventions.

Furthermore, patients who do not complete the MRI scan report an increase in fear, with that fear extending to a 1-month follow-up (Kilborn & Labbé, 1990). In contrast, those completing the MRI scan report a decrease in fear at follow-up (Harris, Robinson, & Menzies, 1999; Kilborn & Labbé, 1990). Kilborn and Labbé found that prescan fear was correlated with postscan claustrophobia, and having to stop the scan was a significant predictor for the subsequent development of claustrophobia. According to Meléndez and McCrank (1993), these results suggest that prescan fear assessment may help predict and allow for interventions to preclude or minimize anxiety-related reactions during and after MRI scans.

Women had higher anxiety ratings both before and after scanning (Mackenzie, Sims, Owens, & Dixon, 1995). Mackenzie et al. evaluated patients' (age = 17–86 years) reactions to MRIs. Approximately 15% ($n = 77$) found a part of the MRI procedure unpleasant. These patients had higher prescanning anxiety scores than those of the other patients, and the differences (as compared with those of the other patients) significantly increased at postscanning. The most unpleasant feature cited by these 77 patients was symptoms of claustrophobia ($n = 19$).

Mackenzie et al. (1995) asked respondents to rate specific MRI features, and those ratings were grouped as "fairly or extremely pleasant," "neither pleasant or unpleasant," or "fairly or extremely unpleasant." Over 20% of the patients rated the following features as being "fairly or extremely unpleasant": (1) the confined space (38%, $n = 175$), (2) the noise of the machine (34%, $n = 150$), (3) lying still during the scan (28%, $n = 128$), (4) moving into the machine (23%, $n = 103$), and (5) being in the scanner alone (22%, $n = 100$). Although 99% of the scans were successful, 10% ($n = 50$) of the patients had State–Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970) anxiety ratings greater than 40 when leaving the unit (Mackenzie et al. noted that elective abdominal surgery had yielded patient anxiety ratings of 40 to 43 on the same scale). Of particular note to the present article, 24% ($n = 120$) of the patients said they would have another MRI, "but only if necessary." Thus, even though 99% of the scans were successful, a sizable percentage of the patients found critical MRI features (e.g., confining space and noise) to be quite unpleasant, producing anxiety and fear; and in terms of potential patient participation in follow-up MRI experiences, many of the patients indicated that they were unwilling to participate, unless it was absolutely necessary.

Both the Cognitions Questionnaire (COGS; McIsaac, Thordarson, Shafran, Rachman, & Poole, 1998)—a measure of patients' thoughts of their experiences in an MRI—and a separate measure of patients' feelings of claustrophobia in an MRI significantly correlate with a subjective rating of MRI-related anxiety; the anxiety scale ranged from 1 (*none*) to 5 (*extremely anxious*). Of the 80 patients in the study, 11 (13.8%) panicked during the scan, and 3 of the 11 who panicked terminated the scan prematurely. The panic group had significantly higher scores on the claustrophobia measure at prescan and on both the prescan and postscan COGS scores. Five patients (10% of the follow-up sample receiving postscan questionnaires 1 month later) reported that they would never go into an MRI again, and another 11 (23%) were only slightly willing to have another scan. Thus, a third of this sample of 17- to 82-year-olds were at least extremely reluctant to undergo an MRI again.

Employing Lukins et al.'s (1997) fear survey schedule for MRIs (FSS–MRI scale), which includes nine of the items from the original FSS scale of everyday fears (Wolpe & Lang, 1964), Harris and colleagues (Harris, Cumming, & Menzies, 2004; Harris, Robinson, & Menzies, 2001) found that, of the different anxiety and claustrophobic measures administered to patients a week prior to their experiencing an MRI, the FSS–MRI scale was the strongest and most consistent predictor of panic symptoms (e.g., sweating) and subjective anxiety in an MRI. Panic and subjective anxiety in the MRI were assessed immediately after the scan. The nine items of the FSS–MRI are the noise of vacuum cleaners, loud noises, thunder, sirens, sudden noises, being in an elevator, being in enclosed spaces, being in airplanes, and being alone. Harris and colleagues noted that the anxiety experienced in the MRI had a reciprocating effect with FSS–MRI; that is, FSS–MRI scores increased between prescan and postscan among those patients who experienced significant anxiety during the scan.

In studies in which there is one MRI scanning session (typically cross-sectional designs), the FSS–MRI scale (Harris et al., 2004; Harris et al., 2001) and other corollary measures, such as the observation behavioral distress composite and three of its items (Elliott, Jay, & Woody, 1987), especially the anxiety-sensitive items that Elliott et al. dropped and that were, therefore, excluded from subsequent studies (e.g., Tyc, Fairclough, Fletcher, Leigh, & Mulhern, 1995), may be used to prescreen subjects who will be undergoing MRIs. One may also use acclimation procedures (Davidson, Thomas, & Casey, 2003; Epstein et al., 2007; Rosenberg et al., 1997; Slifer, Koontz, & Cataldo, 2002; Tyc, Leigh, Mulhern, Srivastava, & Bruce, 1997), such as employing simulators or rewards (e.g., trophies for bravery to decrease subjects' adverse reactions to MRIs), with a hypothesized subsequent decrease in attrition. One may employ the FSS–MRI scale (and/or other measures of distress) as initial measures of MRI anxiety and then relate the findings to simulator acclimation procedures that minimize adverse reactions and dropout in either an actual MRI or a simulator. It is important to note that the FSS–MRI literature has focused on adults, whereas desensitization techniques (Rosenberg et al., 1997), cognitive behavioral techniques (Jay, Elliott, Katz, & Siegel, 1987; Tyc et al., 1997), and operant-contingency techniques (Epstein et al., 2007; Slifer et al., 2002) have

focused on children. Different items, norms, and desensitization protocols may be needed for younger populations. Furthermore, Bangard et al. (2007) found that 33 out of 36 patients (91.7%) who were both anxious and claustrophobic as determined by the Spielberger State–Trait Anxiety Inventory, the FSS–MRI, and a claustrophobia questionnaire completed their MRI scan in an open MRI, as opposed to their prior experience in a closed MRI when 15 out of 36 (41.7%) completed their scan.

Operant conditioning paradigms have been shown to decrease head movements in MRI simulators, showing that these artifactual movements could possibly subsequently decrease during actual MRI sessions (Epstein et al., 2007; Slifer et al., 2002). Epstein and colleagues acknowledged that they did not find differences between their ADHD children and controls because of the low power of the design of their study (e.g., high within-group variability). As the present article and Czarnolewski (2005), point out, such design and statistical issues are more likely to occur for ADHD subjects, for example, than for control subjects.

Given the variances among both different ages and types of children in their acclimation rates and anxiety reactions when they undergo MRI procedures, one would expect to find differences in the likelihood that researchers would be successful in obtaining reliable scans from children with these anxiety-sensitive conditions. For example, Castellanos et al. (2002), who compared children with ADHD to controls, had a large sample ($N = 296$) but did not report refusal rates to initial or subsequent MRIs as part of their longitudinal study. However, they did acknowledge that subjects with ADHD who completed scans ($n = 34$, 11%) were more likely to have blurred scans than normal controls who completed scans were ($n = 16$, 6%). The presence of a differential likelihood of blurred scans was a readily apparent effect of the MRI experience. Subsequent analyses of these data showed an overall attrition rate of 35% for the ADHD group (Shaw et al., 2006). Given the literature reporting that subjects differ in their willingness to retake MRIs, one could logically expect to find an impact on participation rates in longitudinal studies, and the attrition rate in the Shaw et al. study is consistent with this expectation.

Behavioral and Medical Research Models of Validity

Given that these behaviors hypothetically describe subjects who are more susceptible to aversive stimuli, one can evaluate the effects of the MRI measurement process as one would evaluate the effects of an intervention with a number of potential or likely confounds. That is, given the likely impact of experimental confounds on the research design of studies employing MRIs, it appears appropriate to delineate relevant experimental confounds from the research design literature and to suggest their likely presence in MRI studies. First, this section will discuss the relevant literature from the behavioral sciences (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002) and its delineation of threats to internal and external validity. In general, threats to internal validity focus on experimental design issues within the experiment, and threats to external validity focus on experimental design issues affecting the generalizability of the experiment's results. The present article will also delineate relevant research design issues from the medical and epidemiological literatures, the latter being discussed in terms of randomized (Moher, Schulz, & Altman, 2001) and nonrandomized (Des Jarlais, Lyles, Crepaz, & the TREND Group, 2004) studies.

There are a number of potential threats to internal and external validities (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish et al., 2002) that appear relevant to MRI studies. These threats have an impact on the statistical models employed for analyzing data gathered in these studies and would thus be subject to threats of statistical conclusion validity (Cook & Campbell, 1979; Shadish et al., 2002)—that is, a threat to the validity of a statistical model used to evaluate the observed correlations among the data. At issue is whether effects of the MRI experience, such as anxiety reactions, likely cause differential reactions that introduce confounds into the observations obtained from the MRI. Of particular note is the internal validity literature's delineation of the confounds associated with instrumentation effects, which address the MRI setting, and with testing effects, which address subject reactions in that setting.

There are critical differences between the instrumentation and testing threats to internal validity. Instrumentation differs from testing because the former involves a change in the instrument and the latter focuses on a change in the subject (Shadish et al., 2002). On the one hand, the testing confound can be described in terms of the two kinds of reactions represented by (1) subjects who can tolerate and/or eventually can learn to tolerate the MRI experience and (2) subjects who cannot tolerate and/or do not learn to tolerate the MRI experience. Of note, image quality is determined by scan time, which is influenced by the patient's (subject's) willingness to be still (Symms, Jäger, Schmierer, & Yousry, 2004). For longitudinal studies, this dichotomy can be extended by conceptualizing two sets of subject reactions; for example, there are subjects who can tolerate or eventually learn to tolerate the MRI during the first session, and there are those who do not tolerate the MRI during the first session in the MRI. The subjects may be further characterized in terms of their consistency in reacting to the MRI experience. The implication of a behavioral consistency in tolerating the MRI or of lack of behavioral consistency in tolerating the MRI will be discussed later. For now, the point is that when we consider testing confounds, we are focusing on the differential reactions found among subjects.

In contrast to the testing confound, the classical conceptualization of the instrumentation confound focuses on changes in the instrument. In the MRI setting, this may involve placing subjects who tolerate the MRI environment in a closed MRI (with an opening at one end of the machine only) and placing subjects who cannot tolerate the enclosed environment in various types of open MRIs (Hushek et al., 2008; Jouandet, 2003; Spouse & Gedroyc, 2000). The instrumentation confound between the closed and open MRI environments may be shown by the difference in scan resolution between closed and open MRIs (Jarrett, 2000). Procedural changes, such as lengthening

the time to allow subjects to become more acclimated to the MRI before the actual scanning phase of the MRI session, do not result in differences in the resolution of the MRI machine itself. However, procedural differences that include, for example, using dye for some subjects (where permitted by IRB review), which would increase MRI resolution, may not be present when other subjects who are members of a diagnosed group have allergies or refuse to get the injections for administering the dye, thus resulting in an instrumentation confound. Again, the dye increases the resolution, so a difference in usage introduces an instrumentation confound. However, administering dye can introduce testing confounds as well, because we are dealing with subject reactions in their allowance of subjecting themselves to the administration of dye in medical procedures (Elliott et al., 1987; Tyc, Klosky, Kronenberg, de Armendi, & Merchant, 2002) and in MRI procedures, in particular (Tyc et al., 1997).

Another example of an instrumentation effect would be when the experimenter switches to open MRIs (Jouandet, 2003; Spouse & Gedroyc, 2000) during the experiment. The instrumentation effect becomes a factor within the experiment. It also becomes a factor when comparing effects across experiments, such as in a meta-analysis, if there are MRI machines associated with different dropout rates.

The present article proposes a more encompassing perspective than instrumentation and testing confounds when trying to articulate confounds that are inherent in the MRI experience. Because the MRI literature describes the MRI experience as an overwhelming or stressful experience for a sizable proportion of subjects and identifies subject characteristics, such as demographic correlates of age and gender, and the emotional reactions as measured by reliable rating scales, the present author suggests considering the confound introduced by the MRI experience as a *measurement process* confound, which is a more generic and encompassing confound. Such a conceptualization is intended to capture the reactions and corresponding hurdles that many subjects have to overcome in order to endure the measurement process, which allows for a reliable scan that is not blurred due to their possible physical activity (e.g., fidgeting) in reaction to the MRI. The present author notes that the instrument is not changing, per se, when the subject is scanned (i.e., an instrumentation effect); nor is the result of what the subject learns, in terms of tolerating the environment, a testing effect, in which there would be a change in the dependent variable of brain volume. (Again, the classical instances of testing are when the subject learns from the instrument the correct answer for subsequent testing or the subject provides a different emotional response that affects the subject's subsequent rating along a scale.) The confound is neither instrumentation nor testing, but a measurement process that has corresponding predictors (e.g., demographic variables, anxiety measures) of unreliable scans and attrition, which are inherent in and a result of that process.

Attrition

Research design texts (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish et al., 2002) describe attrition (sometimes called *experimental mortality*) as subjects' failure to complete an experiment's outcome measures. If different kinds of people remain to be measured in one condition versus another, such differences could produce posttest outcome differences, even in the absence of treatment (Shadish et al., 2002).

Cook and Campbell (1979) noted that interactions of selection and treatment limit generalization from the experiment when there are selection biases in an experiment. One way to reduce these biases, they said, is to make cooperation in the experiment as convenient as possible. Approaches for increasing participation and decreasing the threats to external validity may prove successful (Rosenberg et al., 1997). Briefly, Rosenberg et al. showed that subject practice in an MRI simulator decreased subject anxiety about being placed in a real MRI machine. Operant conditioning paradigms have been shown to decrease head movements in MRI simulators, thus providing the possibility that these artifactual movements would subsequently decrease during actual MRI sessions (Epstein et al., 2007; Slifer et al., 2002). Thus, one can increase participation rates in an initial MRI session and, given the increased likelihood of, at least, a less-than-aversive experience, potentially increase participation in a follow-up session. This, in turn, increases the generalizability of the results, thereby increasing the study's external validity.

Selection by instrumentation is another critical confound. Historically, this label has been applied when different groups score at different mean positions on a test whose intervals are not equal. Examples are ceiling effects and floor effects. In terms of the MRI, the present author suggests that different likelihoods of participation for each group may artificially create group differences in floor effects (e.g., measures of healthy behaviors) or in ceiling effects (e.g., measures of dysfunctional behaviors).

Shadish et al. (2002) noted that, at times, attrition is caused by the research process. The demands of research exceed those normally expected by treatment recipients. An example is the trade-off between the researcher's desire to measure many relevant constructs as accurately as possible and the respondent's desire to minimize the time spent answering questionnaires or to minimize the time spent in the MRI. They also note that *measurement attrition*, a recurring observation in the MRI literature, refers to a failure to complete outcome measurements, whether or not treatment is completed. *Treatment attrition* refers to cases in which research participants do not continue the treatment, whether or not they continue taking the measurement protocol. If different kinds of people remain to be measured in one condition versus another, then such differences could produce posttest outcome differences, even in the absence of treatment. Attrition is therefore a special subset of selection bias occurring after the treatment is in place. But unlike selection, differential attrition is not controlled by random assignment to conditions.

Looking at things from the front end of MRI studies, Campbell and Stanley (1966) stressed that the selection $\times$ treatment interaction contains the possibility that the effects validly demonstrated hold only for that unique population from whom the experimental and control groups

were jointly selected. This situation becomes more likely when we have more difficulty getting subjects for an experiment. These authors' historical examples consider problems resulting from the greater amount of cooperation required between the subject, school, and so on and the experimenter. Campbell and Stanley's general rule is that the greater the amount of disruption of routine, and the higher the subject refusal rate, the more opportunity there is for a selection-specificity effect. The MRI literature cited in the present article shows that these effects are likely when MRIs are used.

Shadish et al. (2002) noted that reviews of the literature have suggested strategies for reducing measurement attrition that are specific to the kinds of outcome measures used. For example, the "using the foot in the door" method gets the respondent to agree to a smaller task first and a larger one later. A similar rationale may be applied to using an MRI simulator to get a subject to use the real MRI. Davidson et al. (2003), Epstein et al. (2007), and Rosenberg et al. (1997) are examples where this approach is found in the MRI literature.

Shadish et al. (2002) noted the necessity of documenting different aspects of attrition. These include reporting the attrition rates, overall and between groups, and identifying different patterns of attrition, whether corresponding to different measurements or to different types of subjects. They also noted that it is necessary to estimate effects in the data sets that are due to attrition by analyzing data with and without imputed data, and they further suggested the use of multigroup structural equation models (S.E.M.s) for estimating effects in the presence of missing data (this will be discussed in the Statistical Models and Missing Data section). They noted that small sample sizes do not have stable parameters with S.E.M.s. In general, Shadish et al. (2002) suggested modeling the drop-off process itself. The present article's citation of variables in the MRI literature that are potential targets for modeling attrition in the data include gender, age, anxiety tolerance measures, or measures of anxiety tolerance in different environments, such as the FSS, and measures of claustrophobia.

Medical and Epidemiological Standards

Since the MRI was developed and initially used in medical studies, it is critical to highlight that medical research models of validity would also be relevant for evaluating MRI studies. These approaches to validation are exemplified by the consolidated standards of reporting trials (CONSORT) for random clinical trials (Moher et al., 2001) and the transparent reporting of evaluations with nonrandomized designs (TREND) for nonrandom studies (Des Jarlais et al., 2004), both of which are becoming standard when reporting large-scale studies.

The CONSORT statement is a checklist of critical aspects of clinical trials that authors are increasingly being required to report in the manuscripts they submit to journals (Moher et al., 2001). Likewise, the TREND statement is a checklist for verifying that standards of behavioral and public health intervention evaluations involving nonrandomized designs are followed. The checklist is meant to be consistent with the CONSORT checklist and contains items relevant to behavioral and public health intervention studies, whether or not randomized designs are used (Des Jarlais et al., 2004).

The CONSORT and TREND lists contain recommendations for reporting details that may have an impact on the threats to internal and external validity of an experiment. One primary concern motivating the CONSORT recommendations, which may also apply to MRI studies, is that participants who were excluded after allocation to a treatment are unlikely to be representative of all of the participants in the study. The MRI literature shows that patients may experience an aversive situation in MRI studies, regardless of whether they are administered a treatment. In terms of the CONSORT focus, patients may not be available for follow-up evaluation because they had experienced an acute exacerbation of their illness or severe side effects of treatment (Altman et al., 2001).

A critical innovation of the CONSORT standards is the employment of flowcharts to document subject refusal at different points of assessment and treatment. The suggested flowchart of subject assessment and experience recommends reporting the number of subjects excluded for the following reasons: (1) not meeting inclusion criteria, (2) refusing to participate, or (3) other reasons. Furthermore, there should be subsequent reporting of the number of subjects in each group for (1) the allocation phase (*n*s for those who did or did not receive intervention, with reasons); (2) the follow-up phase (*n*s for those who were lost to follow-up, discontinued intervention, with reasons); and (3) the analysis phase (*n*s excluded from analysis, with reasons) (Egger, Jüni, & Bartlett, 2001). For each of the phases, Altman et al. (2001) and Egger et al. recommended detailing the *n*s included, the *n*s not included or excluded, and the corresponding rationale for inclusion and exclusion for each group. For treatment sessions over time, *n*s for each group are reported, as well as *n*s for adverse effects due to treatment.

Adapting these criteria of identifying the *n*s for each phase of the MRI measurement process may include reporting the *n*s who refuse to participate when placed in an MRI; reporting the *n*s who start the MRI, but do not complete the scanning process; and reporting the *n*s who complete the scanning process, but have blurred MRIs because of excessive movement. For MRI studies in which interventions (e.g., medications or cognitive behavioral therapy) are administered, general principles for reporting patient inclusion/exclusion may be adapted to MRI studies. Potential reasons for inclusion/exclusion may include the following: The patient was found not to have the psychiatric/psychological condition; the identified condition required alternative intervention; the intervention was attempted, but it failed or caused an adverse effect; or an alternative intervention was applied because it was better suited to that condition (Altman et al., 2001). Altman et al. used these categories for evaluating cardiovascular studies, such as those reporting not using an intervention (e.g., not implanting a stent because the patient did not require one or was referred to an alternative intervention).

The TREND checklist was based on the fact that evaluation studies very often do not qualify as randomized trials;

therefore, the TREND checklist focuses on methods that equate subjects at baseline and stresses the need to document baseline data and ensure baseline equivalence. Documenting baseline data includes reporting demographic characteristics of each group, which is part of reporting baseline characteristics for each study condition relevant to specific disease prevention. It also includes reporting behaviors, baseline comparisons of those lost to follow-up, and those retained—both overall and by study condition.

Baseline equivalence focuses on studying group equivalence at baseline and the statistical methods used to control for baseline differences. A primary concern is the lack of generalizability resulting from adverse events. Generalizability, according to TREND, requires taking the following into account: the study population; the characteristics of the intervention; the length of follow-up; the incentives, compliance rates, and specific sites/settings involved in the study; and other contextual issues. It is fair to state that, very often, MRI studies are not randomized and would be subject to these data-reporting requirements, which are similar to CONSORT requirements for reporting. Another requirement similar to CONSORT requirements would be reporting the attrition rates of subjects at each phase in the MRI measurement process.

Thus, we see that the threats to validity as delineated in both the behavioral and medical science literatures are relevant to studies employing MRIs. The concerns include threats to internal validity, such as testing, instrumentation, and selection by instrumentation effects (or, more generically, selection by measurement process effects), and threats to external validity, which are primarily due to the attrition and differential attrition effects. Statistical conclusion-validity issues will be expanded upon in the Statistical Models and Missing Data section. Employing an example from the literature will help clarify how these effects likely impact the interpretation of MRI studies.

An Example From the Literature

So far, the present article has presented findings from the literature—namely, the pattern of differential attrition rates that result in threats to internal and external validities in studies that employ MRIs. It has also described the testing and instrumentation confounds relevant to validity, as represented by both the behavioral science and medical literatures, and subsumed the testing and instrumentation confounds under a measurement process confound. We can now provide an example from the literature, the results of which could have been affected by these confounds. The example will show how it is possible to obtain biased sampling, which, due to the selective loss of subjects when using MRIs, results in threats to internal and external validity.

However, before discussing whether the MRI measurement process introduces confounds due to its representing a stress-laden environment, we first have to ask whether anxiety, per se, is represented in or related to biological substrates, such as morphological measurements of brain size and other measures. The answer is yes. Reductions in left amygdala gray matter volume were noted for pediatric patients with anxiety disorders, as compared with comparison subjects (Milham et al., 2005). State–trait anxiety scores were related to fMRI measurement of basolateral amygdala activation in response to backwardly masked, threat-related stimuli (Etkin et al., 2004). When measured by magnetic resonance spectroscopy (with an MRI machine), *N*-acetylaspartate was more concentrated in the left hemisphere orbital frontal cortices in 8 healthy subjects with state–trait anxiety scores higher than 70 than in 8 subjects with scores below 70 (Grachev & Apkarian, 2000). The anxiety scores were determined right before the subjects entered the MRI.

In a meta-analysis of studies employing fMRIs, it was reported that amygdalar and insular hyperactivity were present in subjects with PTSD, social anxiety, and specific phobias, as well as during fear conditioning in healthy subjects (Etkin & Wager, 2007). Etkin and Wager further discussed a common pathway for anxiety reactions. Thus, different methodologies allow for corroborative evidence that supports the position that individual differences in biological substrates are related to individual differences in states of anxiety. Given these results, it is reasonable to investigate other results in the literature in which subjects are exposed to an MRI and to hypothesize whether the pattern of results of reported biological substrate differences were confounded by the MRI measurement process, which has been shown to elicit anxiety reactions and corresponding differential attrition rates.

A pattern of results in Sowell et al. (2003) may indicate an MRI screening effect, especially in ADHD studies. Sowell et al. reported pronounced structural brain-size differences between children with ADHD and controls, as well as a limited number of strong correlations between these measures and the behavioral measures in the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994) for diagnosing the presence of ADHD. Sowell et al. acknowledged that their sample size precluded differentiating among ADHD subtypes, including those with comorbid anxiety disorder, which, with mood and disruptive behaviors, are found to have higher rates among children with ADHD (Busch et al., 2002).

The children were selected from a larger group of children with ADHD whom Sowell et al. (2003) had previously studied, including 11 girls who ranged from 8 to 16 years of age ($M = 11.6$, $SD = 2.8$) and 16 boys who ranged from 8 to 18 years of age ($M = 12.8$, $SD = 3.2$). There were also 17 control girls with a mean age of 11.8 ($SD = 3.1$) and 29 control boys with a mean age of 12.2 ($SD = 3.3$) (age ranges were not reported for controls). There were no gender effects, so the experimenters combined the boys' and girls' data within each group. Sowell et al. reported reduced brain-surface extents (distance from the center of the brain) in the lateral aspects of the anterior temporal cortices and in the inferior portion of the dorsolateral prefrontal cortices for children with ADHD, as compared with those for control children. They also reported more concentration in gray matter (gray matter density) in the bilateral parietal and temporal lobes of the ADHD group, as well as in the right occipital lobe region.

Sowell et al. (2003) reported two striking structural size and behavioral measure correlations. Within the ADHD

group, the correlation between gray matter density and inattention measures in the left occipital lobe showed that the smaller the gray matter density, the more inattentive the patient (they did not report the correlation, but, assuming $n = 27$, it would have to have been about .55 to produce $p < .003$, two-tailed). They also showed that the greater the brain-surface extent in the mesial dorsal frontal region, the higher the hyperactivity score within the ADHD group ($r = .51$, $p < .009$). Similar but less dramatic correlations are found for the larger regions of the dorsal frontal region in the right ($p < .049$) and left ($p < .065$) hemispheres.

There appear to be relationships between Sowell et al.'s (2003) group mean differences and correlations. First, regarding the frontal gray areas, the larger the (mesial) dorsal frontal region (i.e., the greater the brain-surface extent), the higher the hyperactivity score. Therefore, those ADHD subjects at the upper end of the hyperactivity scores would have been less likely to be included because, as these data show, they were more likely to be too active or fidgety to remain still in the stressful environment of the MRI. One possible interpretation is that the most hyperactive children with ADHD (i.e., those who scored at the extreme high end of the scale) were less likely to be included because, all other things being equal, the hyperactive children would be less likely to lie still in the stressful MRI environment. Other ADHD subjects possibly less likely to be included were those who scored at the upper end of, but within the same region of, the scale as the ADHD children who were included; but, unlike the children with ADHD who had similar hyperactivity scores and did manage to tolerate the stressful MRI environment, the excluded children could not tolerate the stressful MRI environment.

Thus, one implication of the positive correlation between hyperactivity and the size of the mesial dorsal frontal region ($r = .51$) for the children with ADHD is that, given that their reported average size difference was *larger* than that found for controls, the difference in size of the frontal regions between children with ADHD and controls may actually have been *larger* than that reported. We are assuming a positive correlation in sizes for most frontal areas within the ADHD (and within the control) groups. Similarly, the greater similarity of the other frontal regions either in location or function as the mesial frontal region, the *more* likely that the reported ADHD frontal region sizes were *larger* as well, and, therefore, the *more* likely the reported frontal region differences between ADHD subjects and controls were *larger*. We are not even considering the issue of a restricted range of scores, a threat to statistical conclusion validity (Cook & Campbell, 1979; Shadish et al., 2002) that would mean that the hyperactivity score and frontal lobe size relationship is stronger, thus adding further emphasis to the points just made.

A similar argument may be made regarding the gray matter density differences reported for the occipital lobe. Again, the *smaller* the gray matter density of the left occipital lobe, the *higher* the inattentiveness score within the ADHD group. It is also reasonable to assume that the occipital regions within the ADHD group (and within the control group) correlated positively with each other. Furthermore, one may reasonably hypothesize that the ADHD subjects at the upper end of the inattentiveness scale were less likely to be included in the study because they would be less likely to focus on cognitive coping strategies (see, e.g., Epstein et al., 2007; Rosenberg et al., 1997; Tyc et al., 1995) that would enable them to remain sufficiently still in the stressful environment of the MRI. ADHD subjects at the upper end of the scale would be less likely to be included, as well as the ADHD subjects who scored at the upper end (as opposed to the extreme end) of the scale but could not tolerate the stressful MRI environment. Therefore, we have a situation where ADHD subjects having *lower* gray matter density would be *more likely* to experience attrition, which would result in an artifactually *greater* gray matter density for the ADHD group. Again, one may assume that the occipital lobe gray matter densities within each group are positively related. Thus, given that ADHD subjects are reported to have more gray matter density in the right occipital lobe region, the reported occipital lobe differences between the ADHD and control groups are likely to be artifactually *larger* than reported. Therefore, the reported differences between the ADHD group and controls should be *smaller*.

Sowell et al.'s (2003) findings suggest a confound that may be present in longitudinal studies (Castellanos et al., 2002). For example, the frontal lobe distance from the center of the brain is positively related to the hyperactivity scores of children with ADHD in the MRI experiment, thus resulting in artifactually smaller frontal lobes. Growth from an artifactually smaller ADHD distance from center, all other things being equal, would result in artifactually larger frontal lobe growth than would be reported for children with ADHD.

There is, however, a reasonable alternative argument. Castellanos et al. (2002), besides their other criteria, selected children with ADHD who had Conners' Teacher Rating Scale hyperactivity ratings more than 2 *SD*s above the mean ratings for comparable age- and gender-specific groups. This likely resulted in the selection of the ADHD children who had larger gray matter frontal lobes, given Sowell et al.'s (2003) data. However, Castellanos et al. reported, on average, smaller frontal lobe gray matter at baseline than was found for their controls. Given the reported 35% attrition rate in this data set for children with ADHD (Shaw et al., 2006), it was likely that the restriction range in high hyperactivity scores would result in a corresponding expectation of disproportionately higher attrition rates in the stressful MRI environment. Given Sowell et al.'s data, this would have resulted in a correspondingly disproportionate loss of subjects with larger frontal lobes at the follow-up scans. Thus, children with ADHD would show smaller frontal lobe growth rates than controls would. The lack of expected reliable differences in frontal lobe size growth trajectories might have been a function of selection criteria that increased the likelihood of differential attrition rates in the stressful MRI environment for the more hyperactive ADHD subjects.

A curvilinear (negative accelerating) form of cerebral volume change, as a function of age among a large normative group, has also been reported (Giedd et al., 1999). Therefore, the age variable, which also relates to cerebral

region volume size, may likely have an impact on the possibility of finding differences in volume size, as well as on the subsequent differential changes in volumes between children with ADHD and controls.

Gender is another variable showing differential volume change, with girls having maximal volumes at age 11 and boys having maximal volumes at age 12 for both frontal and parietal volumes (Giedd et al., 1999). This issue of the selective loss of subjects, which is related to the very same variables (age and gender) that the MRI literature showed to be related to the likelihood of aversive reactions to MRIs, would further confound brain size comparisons between ADHD subjects and controls. This would occur when making comparisons across age and gender for each of these groups and when observing these groups in longitudinal studies. Given the subject reactions cited in the MRI literature, these suggested confounds appear even more likely to occur in longitudinal MRI studies in which there are children with comorbid anxiety predispositions, such as children with ADHD. Again, Castellanos et al. (2002) reported curvilinear trajectories for children with ADHD, but they also reported minimal differences between ADHD subjects and controls, except with regard to the caudate nucleus. At issue is whether their findings of a lack of differential changes may have been due to the confounds articulated above.

Using the MRI literature on subject reaction and using the attrition rates found in the Castellanos et al. (2002) data set (Shaw et al., 2006), we can develop statistical models for differential attrition rates for gender, age, and gender × age interactions (Shadish et al., 2002). The purpose of this modeling would be to compare the attrition rate trajectories with the trajectories reported in Castellanos et al., with the further possibility of incorporating these attrition rates into the model employed by Castellanos et al. for testing differences in trajectories. (The author thanks a reviewer for this critical point.)

In terms of the vocabulary used in the literature about threats to internal and external validity, the supposed differential attrition rates in MRI studies are likely to produce an experimental confound defined by a selection × measurement process interaction (Campbell & Stanley, 1966, used the term *selection × treatment interaction*; Cook & Campbell, 1979; Shadish et al., 2002). The present article considers MRI trajectory growth studies as observational studies and considers as well the studies having an inherent selection × measurement process interaction, which is part of the MRI measurement environment. With the selection × measurement process interaction (per Campbell & Stanley, 1966), different groups score at different mean positions on a test whose intervals are subject to ceiling effects in which a subsequent measurement is limited in getting higher because the initial measurement, on average, is large/high (and larger than the comparison group). The selection × measurement process interaction may also have floor effects in which subsequent measurement is limited in getting lower because the initial measurement, on average, is small/low (and smaller than the comparison group).

The present author further suggests that the differences in likelihood of group participation in MRI studies may be a reflection of group differences in resilience in tolerating the stressful MRI environment. This is consistent with Etkin and Wager (2007), who stated that frontal lobes exert a modulating effect on brain regions involved in anxiety reactions. The stressful environment may artificially create group differences in either floor or ceiling effects of the MRI volumes for patients who are more likely to be in a stress-sensitive group, which would logically include ADHD subjects. These effects would be compounded in longitudinal studies for the ADHD group. Again, the ADHD subjects would more likely be hyperactive and, therefore, would more likely be subject to attrition with resultant artifactually smaller frontal lobe size (on average) and inflated growth. This artifactually inflated growth may obfuscate differences in growth rates between the ADHD and control groups.

Such a scenario suggests that Sowell et al.'s (2003) finding that the control group has, on average, a smaller distance from the center of the brain for the dorsal frontal lobe may be understated. The lack of ADHD–control group differences in growth rates in the frontal lobes (Castellanos et al., 2002) appears to be inconsistent with both these authors' hypotheses—and others' (Berger & Posner, 2000)—that the frontal lobes are the seat of executive functioning and represent a primary function that distinguishes children with ADHD from controls.

Statistical Models and Missing Data

A primary focus of the present article is the significant impact that a specific method of measurement—in this case, the MRI—has on the likelihood of obtaining successful measurements in a certain group of subjects. This is related to and has a significant impact on which anatomical or morphometric makeups of brains are more likely to be successfully scanned in an experiment, with the researchers' very hypotheses being focused on associating morphological differences of subjects who differ in, for example, their reactions to anxiety-inducing situations. Confounds that may be introduced because of the aversive environment in the MRI measurement procedure need to be minimized in order to minimize the threats to internal and external validity.

We will now discuss statistical models (Allison, 2003; Nich & Carroll, 2002; Shadish, Luellen, & Clark, 2006) that delineate the factors that have an impact on the presence of missing data. It is necessary to point out that the present article does not consider the impact of treatment (i.e., independent variables, such as medications or behavioral therapies) on the pattern of missing data (i.e., missingness) found in a study. Rather, we consider the impact of the dependent variable (e.g., brain volume, as measured by the MRI machine) on the missingness found in a study and its consequent impact on a study's results.

Again, Shadish et al. (2002) provided a detailed discussion distinguishing attrition due to measurement from attrition due to treatment. They noted that the distinction is practically important for several reasons. First, measurement attrition prevents the inclusion of the participant in the analysis (except via missing data imputation methods), but treatment attrition does not preclude inclusion, as long as the participant completes the measures. Second, many dropouts can be convinced to complete the measurement

protocol, even when they refuse to complete the treatment. Third, if measurement attrition can be eliminated, the researcher can implement a classic intent-to-treat analysis, and, if a good implementation measure is available, the researcher can sometimes use certain analytic strategies. Shadish et al. (2002) stressed a good rule: Prevent measurement attrition, even when you cannot prevent treatment attrition.

For clinical trials with random assignment to treatment, CONSORT stated that "an intent-to-treat design requires that all patients are followed according to the pre-specified schedule with principal, and perhaps secondary, outcome assessments regardless of compliance, adverse effects, or other post-randomization, observations—death and patient refusal excepted" (Lachin, 2000, p. 183).

Intention-to-treatment data are analyzed with regard to the group to which participants are randomized, irrespective of any deviation from the treatment protocol for some of the subjects. When none of the subjects fully participate in an experiment, we have missing data, and the missingness potentially introduces confounds into the study. The literature distinguishes among three types of data sets with missing data (Wright & Sim, 2003).

1. Missing completely at random (MCAR): The missing observations on a variable constitute a random subset of that variable, and values of the missing observations are not related to other variables.

2. Missing at random (MAR): The missing observations on a variable constitute a random subset of that variable, but the values of the missing observations are related to those of another variable.

3. Missing nonrandomly (MNAR): As stated by Wright and Sim (2003),

> The missing observations are not a random subset of a variable, and the values of the missing observations may also be related to those of some other variable. . . . Unless unobserved values are missing completely at random, they may lead to a loss of between-group comparability and thus bias in estimation of treatment effects. (p. 834)

According to Schafer and Graham (2002), these definitions may be best understood in the context of modern missing-data procedures, which regard missingness as a probabilistic phenomenon. The distribution of *R*, representing missingness, is regarded as a mathematical device for describing the rates and patterns of missing values and for capturing possible relationships between missingness and the values of the missing items themselves. Complete data are separated into observed and missing data. Schafer and Graham pointed out that MNAR is present when the distribution of missingness depends on the data that is missing from the dependent variable. MCAR is present when missingness is not due to either the independent or dependent variables, and MAR is present when missingness is due to some nonexperimental variables' relationships to the experiment's independent (or covariate) variables.

Alternative statistical models have been proposed for capturing variable effects when there is missingness among the variables. Three data-analytic, potentially complementary approaches are the use of (1) random effect models (Nich & Carroll, 2002), (2) structural equation models (S.E.M.) (Allison, 2003; McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002; McArdle et al., 2004), and (3) propensity scores (Shadish et al., 2006). Each approach potentially allows for a unique contribution to the missingness issue in the MRI research area.

Nich and Carroll (2002) used a random effect regression model on all of the data for the full, intended duration of treatment in a study. They included a covariate for whether the outcome was collected during treatment (i.e., treatment status). This variable, in addition to being a main effect, becomes part of the interaction terms for each of the variables (and combination of those variables) in the study. Treatment status is used as a time-varying covariate; that is, it assesses treatment differences during the intended duration of the protocol, as well as the effect of treatment dropout. This strategy takes into account the point at which protocol deviation occurs (i.e., dropout, noncompliance, or withdrawal). This approach is consistent with the TREND statement's focus on identifying and quantifying the sources of differential compliance rates. Therefore, not only did these authors follow CONSORT's suggestion of reporting the nature of the protocol deviation, but they also described the specific analytic strategy used, as suggested by the TREND recommendations. Allison (2003) detailed alternative imputation procedures and stressed the use of maximum likelihood estimates when employing structural equation models to deal with missing data in repeated measures designs.

Regardless of the model used, the following Nich and Carroll (2002) statement highlights limitations of the imputation approach:

> At a minimum, investigators should report the number of participants in each treatment group who dropped out or for whom data were unavailable for other (withdrawal, noncompliance, loss to follow-up) as well as how their data were handled (method of imputation, how many values imputed, whether analysis accounted for differential retention or exposure to non-study treatments). (p. 130)

Nich and Carroll (2002) stated that the random effects regression model might provide reasonable estimates, as compared with other imputation methods or with the deletion of subjects with missing values. However, they concluded that all methods have bias.

The issue of generalizability is a focus of the TREND initiative because, by definition, it acknowledges the lack of a random sample in the study. The TREND checklist calls for data-analytic models for considering the study population, the characteristics of the intervention, the length of the follow-up, incentives, compliance rates, specific sites/settings involved in the study, and other contextual issues. Measuring patient acclimation and aversive reactions in MRI studies would provide a better understanding of the data that are gathered as part of the TREND initiative and that detail the incentives and compliance rates in the study.

McArdle et al. (2002) provided an example for studying trends when each subject has only two data points. They

systematically compared alternative growth curve solutions for different structural equation models when clustering subjects on the basis of their initial scores. McArdle et al.'s (2004) use of S.E.M. builds on regression models that employ latent scores (underlying unobservable factor scores) to control for the subject's initial age and, therefore, provide reliable estimates for the change from initial age. McArdle et al. (2004) developed this model in order to predict latent difference scores of brain size between two time periods.

McArdle et al.'s (2004) bivariate, dynamic, latent-score difference S.E.M. contains *x* and *y* latent variables. There are latent *x* and *y* change scores, with each dual-change score having its own linear and proportional (nonlinear) parameters. Of particular interest is the potential application of the model for those interested in testing which brain region or regions are the areas driving the resulting changes in brain size. These coupling parameters include a time-dependent effect of a latent *x* variable onto a latent *y* variable and a time-dependent effect of a latent *y* variable onto a latent *x* variable.

McArdle et al. (2004) employed a vector-field display that allows the researcher to interpret the direction and strength of the dynamics of the coupling. The result is a latent-score difference model that can lead to a pictorial representation of complex nonlinear trajectory equations (i.e., nonhomogeneous equations). The present author suggests employing this model for estimating growth trajectories of MRI data and further suggests that the latent *x* variable can be total cerebral volume (TCV), a common covariate in the MRI literature. McArdle et al.'s (2004) model can then test, for instance, whether each particular region or set of regions (latent *y* variable) drives the growth in TCV, and can test whether these dynamic growth models differ between groups (e.g., diagnostic vs. control).

A further note is that McArdle et al. (2004) acknowledged the impact of attrition. They noted that subjects who dropped out during their study were slightly lower in working memory performance at baseline than were those who participated in the follow-up. They attempted to account for this nonrandom attrition by including all longitudinal and cross-sectional data in the models, but they recognized the possible presence of this confound, especially as related to other key variables, such as age selectivity of a cohort.

Shadish et al.'s (2006) use of subjects' scores of propensity to participate in an experiment appears to be applicable to the focus on missingness in the present article. A propensity score is the conditional probability that a person will be in one condition rather than another (e.g., get a treatment rather than be in the control group), given a set of observed covariates. For example, if an equal-probability assignment mechanism is used to assign people to one of two conditions, each person has a 50% probability (i.e., a propensity score of .50) of being in one of the two conditions. Shadish et al. (2006) noted that quasi-experiments do not have random assignments. Quasi-experiments are more practical than randomized experiments, especially when randomized assignment is not feasible or practical. Through the use of propensity scores, Shadish et al. (2006) quantified the probability of missingness in their study.

With quasi-experiments, the true propensity score is not known and must be estimated. The probabilities of receiving treatment (i.e., propensity scores) are likely to vary from .50 and correlate with individual characteristics that influence treatment selection. Covariates may include convenient demographic variables, such as age and gender.

Shadish et al. (2006) used a number of variables as covariates in an ANCOVA or as measures to stratify the data set to equate subjects. Each of these procedures helped develop propensity-to-participate scores. Using these procedures, Shadish et al. (2006) decreased the effect of random assignment versus self-selection assignment. The result was that the difference in measures of two skills (i.e., that between math and vocabulary) in their self-selection tutored group approached the difference between these scores in the random-assignment group. They showed that variations of these procedures can be differentially successful in balancing subject characteristics, with the corresponding effect on the success in diminishing the differences found for random assignment versus self-selection.

The critical point about propensity scores is that they are based on subject characteristics that may be involved in confounding (either by increasing or decreasing) the experimental effects. The MRI literature suggests that propensity scores are available to help account for subject participation rates. Some measures that may be used to compute propensity scores include measures of anxiety, such as an FSS–MRI measure (Harris et al., 2004; Harris et al., 2001); demographic variables; and other rating scales that may be employed as covariates in MRI studies and may be used as screening devices to identify subjects who are likely to require more intensive acclimation to the MRI. For example, one may employ the FSS–MRI scale (and other measures of distress) to obtain initial measures of MRI anxiety and then relate the results to simulator acclimation procedures that minimize adverse reactions and dropout in an MRI simulator and/or actual MRI. Furthermore, subject scores on the FSS–MRI scale (and distress measures) and subject acclimation rates to the MRI environment may be used as covariates for the pattern of missing data (i.e., missingness) in studies with either a one-time MRI scanning session or repeated scanning sessions in longitudinal studies.

Another strategy is to block subjects in longitudinal studies on the basis of their consistency in tolerating the MRI environment, thus preventing across-group anxiety reactions and corresponding attrition from confounding between-group comparisons.

One may hypothesize that a child who acclimates well enough to MRIs that reliable and complete assessments can be made for the first two scanning sessions (even when the sessions may be years apart) provides a behavioral measure of being different in terms of resilience in tolerating the MRI environment from a child who does not acclimate at the first session but does acclimate at the second session; or, alternatively, a child who does not acclimate at the first two sessions, but acclimates at the third session—again, even when there are years between each session. In short, one can "equate" subjects by their abil-

ity to become sufficiently settled to give researchers two reliable scanning sessions within the same number of sessions (i.e., two). The suggestion is to select for analyses subjects who provide two complete and reliable scans in two successive attempts over time, especially in the first two sessions in the experiment. Thus, one is computing change scores among a homogeneous set of subjects, in which there is no confound due to acclimation to the measurement method. One may expand the blocking by having another block for those subjects who tolerated an MRI scan for three or more consecutive sessions, assuming there is a sufficient number of subjects for the block.

We can see how the MRI measurement process has an effect on the research design of an MRI study. These effects, which are detailed in the discussion of validity threats, have an effect on the statistical parameters of the MRI measurement. There is a wealth of models designed to impute missing data of large clinical trials. Clearly, there is much to be gained by integrating McArdle et al.'s (2004), Nich and Carroll's (2002), and Shadish et al.'s (2006) models in order to address missingness when studying growth trajectories in MRI data.

Implications for MRI Studies

Without equating, whether done statistically or by blocking, one hypothetically could weaken and diffuse the effects of age and group factors when computing trajectories for MRI volumes of children (e.g., children with ADHD), especially across a large age range. This differential attrition could logically result in Type I or Type II errors, depending on the group size differences, group variance differences, and the particular model employed (DeShon & Alexander, 1996).

More types of open MRIs are being developed, are becoming more accessible, and are approaching capabilities of closed MRIs (Bangard et al., 2007; Hushek et al., 2008; Jouandet, 2003; Siemens AG, 2004). How anxiety-provoking is the open MRI (one may reasonably hypothesize much less so)? Also of interest is the possibility of the reanalysis of previously published studies that used closed MRIs and did not employ analytic models that could have addressed the potential confound issues raised by the present article. Another possibility is the pooling of data across data sets (where possible) in order to integrate the three data approaches of employing propensity scores with models that incorporate parameters of participation consistency and dynamic structural equating models. The possibility of Type II errors in previously published studies with large *N*s or small *N*s clearly exists. Reanalyzing data with large *N*s or pooling across studies with smaller *N*s in order to identify effects that may have been missed may provide an incentive for such an effort.

Another incentive is related to the possibility that, given new technologies, MRIs may more likely become included in large-scale studies. Developing statistical models to address subject reactions and consequent participation/attrition rates would be beneficial. Further developing the statistical models in random clinical trials—which traditionally have participation/attrition-rate issues—to also consider the participation/attrition dynamics in MRI studies opens other research possibilities. Statistical models that address subject reactions to both treatment interventions and MRIs (of both controls and diagnostic groups) may thus allow for a more comprehensive look at the effects of the interventions in such large-scale trials.

AUTHOR NOTE

The author acknowledges the very helpful comments of Peter Sheridan and Martha Ann Carey, as well as those of the journal's reviewers, whose comments made all the difference. The opinions expressed in this article are those of the author's and not of the U.S. Department of Health and Human Services, where the author is an employee. Correspondence concerning this article should be addressed to M. Y. Czarnolewski, 11231 Columbia Pike, Silver Spring, MD 20901 (e-mail: mczarnol@yahoo.com).

REFERENCES

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, **112**, 545-557.

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine*, **134**, 663-694.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Bangard, C., Paszek, J., Berg, F., Eyl, G., Kessler, J., Lackner, K., & Grossman, A. (2007). MR imaging of claustrophobic patients in an open 1.0T scanner: Motion artifacts and patient acceptability compared with closed bore magnets. *European Journal of Radiology*, **64**, 152-157.

Berger, A., & Posner, M. I. (2000). Pathologies of brain attentional networks. *Neuroscience & Biobehavioral Reviews*, **24**, 3-5.

Busch, B., Biederman, J., Cohen, L. G., Sayer, J. M., Monuteaux, M. C., Mick, E., et al. (2002). Correlates of ADHD among children in pediatric and psychiatric clinics. *Psychiatric Services*, **53**, 1103-1111.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Castellanos, F. X., Lee, P. P., Sharp, W., Jeffries, N. O., Greenstein, D. K., Clasen, L. S., et al. (2002). Developmental trajectories of brain volume abnormalities in children and adolescents with attention-deficit/hyperactivity disorder. *Journal of the American Medical Association*, **288**, 1740-1748.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Chicago: Rand McNally.

Czarnolewski, M. Y. (2005, August). *Statistical and design issues when using MRIs for brain research*. Poster presented at the 113th Annual Convention of the American Psychological Association, Washington, DC. Retrieved from http://psycnet.apa.org/psycextra/524312009-001.pdf.

Davidson, M. C., Thomas, K. M., & Casey, B. J. (2003). Imaging the developing brain with fMRI. *Mental Retardation & Developmental Disabilities Research Reviews*, **9**, 161-167.

DeShon, R. P., & Alexander, R. A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal. *Psychological Methods*, **1**, 261-277.

Des Jarlais, D. C., Lyles, C., Crepaz, N., & the TREND Group (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health*, **94**, 361-365.

Egger, M., Jüni, P., & Bartlett, C. (2001). Value of flow diagrams in reports of randomized controlled trials. *Journal of the American Medical Association*, **285**, 1996-1999.

Elliott, C. H., Jay, S. M., & Woody, P. (1987). An observation scale for measuring children's distress during medical procedures. *Journal of Pediatric Psychology*, **12**, 543-551.

Epstein, J. N., Casey, B. J., Tonev, S. T., Davidson, M., Reiss, A. L., Garrett, A., et al. (2007). Assessment and prevention of head motion during imaging of patients with attention deficit hyperactivity disorder. *Psychiatry Research: Neuroimaging*, **155**, 75-82.

ESHED, I., ALTHOFF, C. E., HAMM, B., & HERMANN, K.-G. A. (2007). Claustrophobia and premature termination of magnetic resonance imaging examinations. *Journal of Magnetic Resonance Imaging*, **26**, 401-404.

ETKIN, A., KLEMENHAGEN, K. C., DUDMAN, J. T., ROGAN, M. T., HEN, R., KANDEL, E. R., & HIRSCH, J. (2004). Individual differences in trait anxiety predict the response of the basolateral amygdala to unconsciously processed fearful faces. *Neuron*, **44**, 1043-1055.

ETKIN, A., & WAGER, T. D. (2007). Functional neuroimaging of anxiety: A meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. *American Journal of Psychiatry*, **164**, 1476-1488.

FITZ, C. R. (1989). *MRI sedation: The search for a magic bullet.* Presented at the 26th Congress of the European Society for Pediatric Radiology, Dublin, Ireland.

GIEDD, J. N., BLUMENTHAL, J., JEFFRIES, N. O., CASTELLANOS, F. X., LIU, H., ZIJDENBOS, A., ET AL. (1999). Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience*, **2**, 861-863.

GRACHEV, I. D., & APKARIAN, A. V. (2000). Anxiety in healthy humans is associated with orbital frontal chemistry. *Molecular Psychiatry*, **5**, 482-488.

HARRIS, L. M., CUMMING, S. R., & MENZIES, R. G. (2004). Predicting anxiety in magnetic resonance imaging scans. *International Journal of Behavioral Medicine*, **11**, 1-7.

HARRIS, L. M., ROBINSON, J., & MENZIES, R. G. (1999). Evidence for fear of restriction and fear of suffocation as components of claustrophobia. *Behaviour Research & Therapy*, **37**, 155-159.

HARRIS, L. M., ROBINSON, J., & MENZIES, R. G. (2001). Predictors of panic symptoms during magnetic resonance imaging scans. *International Journal of Behavioral Medicine*, **8**, 80-87.

HUSHEK, S. G., MARTIN, A. J., STECKNER, M., BOSAK, E., DEBBINS, J., & KUCHARZYK, W. (2008). MR systems for MRI-guided interventions. *Journal of Magnetic Resonance Imaging*, **27**, 253-266.

JARRETT, L. F. (2000). *Open or closed MRI: What's better for you?* Retrieved from www.onewest.net/~ghosttowncandleco/wfy/pdfs/healthwatch.pdf.

JAY, S. M., ELLIOTT, C. H., KATZ, E., & SIEGEL, S. E. (1987). Cognitive–behavioral and pharmacologic interventions for childrens' distress during painful medical procedures. *Journal of Consulting & Clinical Psychology*, **55**, 860-865.

JOUANDET, M. (2003). *Closed and open magnetic resonance imaging.* Ithaca, NY: Cayuga Medical Center. Retrieved from www.cayugamed.org/articles/read.dbm?ID=293.

KILBORN, L. C., & LABBÉ, E. E. (1990). Magnetic resonance imaging scanning procedures: Development of phobic response during scan and at one-month follow-up. *Journal of Behavioral Medicine*, **13**, 391-401.

LACHIN, J. M. (2000). Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials*, **21**, 167-189.

LUKINS, R., DAVAN, I. G. P., & DRUMMOND, P. D. (1997). A cognitive behavioural approach to preventing anxiety during magnetic resonance imaging. *Journal of Behavior Therapy & Experimental Psychiatry*, **28**, 97-104.

MACKENZIE, R., SIMS, C., OWENS, R. G., & DIXON, A. K. (1995). Patients' perceptions of magnetic resonance imaging. *Clinical Radiology*, **50**, 137-143.

MCARDLE, J. J., FERRER-CAJA, E., HAMAGAMI, F., & WOODCOCK, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, **38**, 115-142.

MCARDLE, J. J., HAMAGAMI, F., JONES, K., JOLESZ, F., KIKINIS, R., SPIRO, A., III, & ALBERT, M. S. (2004). Structural modeling of dynamic changes in memory and brain structure using longitudinal data from the normative aging study. *Journals of Gerontology*, **59B**, P294-P304.

MCISAAC, H. K., THORDARSON, D. S., SHAFRAN, R., RACHMAN, S., & POOLE, G. (1998). Claustrophobia and the magnetic resonance imaging procedure. *Journal of Behavioral Medicine*, **21**, 255-268.

MELÉNDEZ, J. C., & MCCRANK, E. (1993). Anxiety-related reactions associated with magnetic resonance imaging examinations. *Journal of the American Medical Association*, **270**, 745-747.

MILHAM, M. P., NUGENT, A. C., DREVETS, W. C., DICKSTEIN, D. S., LEIBENLUFT, E., ERNST, M., ET AL. (2005). Selective reduction in amygdala volume in pediatric anxiety disorders: A voxel-based morphometry investigation. *Biological Psychiatry*, **57**, 961-966.

MOHER, D., SCHULZ, K. F., & ALTMAN, D. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Journal of the American Medical Association*, **285**, 1987-1991.

MURPHY, K. J., & BRUNBERG, J. A. (1997). Adult claustrophobia, anxiety and sedation in MRI. *Magnetic Resonance Imaging*, **15**, 51-54.

NICH, C., & CARROLL, K. M. (2002). "Intention-to-treat" meets "missing data": Implications of alternate strategies for analyzing clinical trials data. *Drug & Alcohol Dependence*, **68**, 121-130.

PEARLSON, G. D., & CALHOUN, V. (2007). Structural and functional magnetic resonance imaging in psychiatric disorders. *Canadian Journal of Psychiatry*, **52**, 158-166.

ROSENBERG, D. R., SWEENEY, J. A., GILLEN, J. S., KIM, J., VARANELLI, M. J., O'HEARN, K. M., ET AL. (1997). Magnetic resonance imaging of children without sedation: Preparation with simulation. *Journal of the American Academy of Child & Adolescent Psychiatry*, **36**, 853-859.

SCHAFER, J. L., & GRAHAM, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, **7**, 147-177.

SHADISH, W. R., COOK, T. D., & CAMPBELL, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

SHADISH, W. R., LUELLEN, J. K., & CLARK, M. H. (2006). Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 143-157). Washington, DC: American Psychological Association.

SHAW, P., LERCH, J., GREENSTEIN, D., SHARP, W., CLASEN, L., EVANS, A., ET AL. (2006). Longitudinal mapping of cortical thickness and clinical outcome in children and adolescents with attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, **63**, 540-549.

SIEMENS AG (2004, July 29). *Siemens introduces first 1.5 Tesla open bore MRI*. Philadelphia: Siemens. Available at http://www.medical.siemens.com/siemens/en_US/rg_marcom_FBAs/files/Press_Releases/2004/PDF/069.04_Espree.pdf.

SLIFER, K. J., KOONTZ, K. L., & CATALDO, M. F. (2002). Operant-contingency-based preparation of children for functional magnetic resonance imaging. *Journal of Applied Behavior Analysis*, **35**, 191-194.

SOWELL, E. R., THOMPSON, P. M., WELCOME, S. E., HENKENIUS, A. L., TOGA, A. W., & PETERSON, B. S. (2003). Cortical abnormalities in children and adolescents with attention-deficit hyperactivity disorder. *Lancet*, **362**, 1699-1707.

SPIELBERGER, C. D., GORSUCH, R. L., & LUSHENE, R. E. (1970). *Manual for the State–Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

SPOUSE, E., & GEDROYC, W. M. (2000). MRI of the claustrophobic patient: Interventionally configured magnets. *British Journal of Radiology*, **73**, 146-151.

SYMMS, M., JÄGER, H. R., SCHMIERER, K., & YOUSRY, T. A. (2004). A review of structural magnetic imaging. *Journal of Neurology, Neurosurgery, & Psychiatry*, **75**, 1235-1244.

TYC, V. L., FAIRCLOUGH, D., FLETCHER, B., LEIGH, L., & MULHERN, R. K. (1995). Children's distress during magnetic resonance imaging procedures. *Children's Health Care*, **24**, 5-19.

TYC, V. L., KLOSKY, J. L., KRONENBERG, M., DE ARMENDI, A. J., & MERCHANT, T. E. (2002). Children's distress in anticipation of radiation therapy procedures. *Children's Health Care*, **31**, 11-27.

TYC, V. L., LEIGH, L., MULHERN, R. K., SRIVASTAVA, D. K., & BRUCE, D. (1997). Evaluation of a cognitive–behavioral intervention for reducing stress in pediatric cancer patients undergoing magnetic resonance imaging procedures. *International Journal of Rehabilitation & Health*, **3**, 267-279.

WOLPE, J., & LANG, P. J. (1964). A fear survey schedule for use in behaviour therapy. *Behaviour Research & Therapy*, **2**, 27-30.

WRIGHT, C. C., & SIM, J. (2003). Intention-to-treat approach to data from randomized controlled trials: A sensitivity analysis. *Journal of Clinical Epidemiology*, **56**, 833-842.

(Manuscript received August 18, 2008;
revision accepted for publication July 14, 2009.)