

Individual Differences Methods for Randomized Experiments

Elliot M. Tucker-Drob
University of Texas at Austin

Experiments allow researchers to randomly vary the key manipulation, the instruments of measurement, and the sequences of the measurements and manipulations across participants. To date, however, the advantages of randomized experiments to manipulate both the aspects of interest and the aspects that threaten internal validity have been primarily used to make inferences about the average causal effect of the experimental manipulation. This article introduces a general framework for analyzing experimental data to make inferences about individual differences in causal effects. Approaches to analyzing the data produced by a number of classical designs and 2 more novel designs are discussed. Simulations highlight the strengths and weaknesses of the data produced by each design with respect to internal validity. Results indicate that, although the data produced by standard designs can be used to produce accurate estimates of average causal effects of experimental manipulations, more elaborate designs are often necessary for accurate inferences with respect to individual differences in causal effects. The methods described here can be diversely applied by researchers interested in determining the extent to which individuals respond differentially to an experimental manipulation or treatment and how differential responsiveness relates to individual participant characteristics.

Keywords: experimental design, validity, causality, Person \times Situation interaction, Aptitude \times Treatment interaction

Supplemental materials: <http://dx.doi.org/10.1037/a0023349.supp>

How can one begin to investigate whether an experimental manipulation or treatment affects some people more than it affects others? In the social and behavioral sciences, questions of how people differ from one another (individual differences) were historically confined to observational designs. However, the value of integrating correlational approaches with randomized controlled experiments is, without a doubt, profound. As was articulated by Cronbach (1975; also see Cronbach, 1957), the most fundamental implication of the existence of individual differences in responses to experimental manipulations or treatments is that “a general statement about a treatment effect is misleading because the effect will come or go depending on the kind of person treated” (p. 119). Of course, understanding the rules that govern how people respond differentially to treatment or manipulation effects not only can alleviate the concern expressed by Cronbach, but can actually help to develop more nuanced and accurate understandings of scientific constructs and psychological processes. Moreover, investigation of individual differences in treatment effects and their correlates can have pragmatic applications, for example, at the individual level

(a) helping to choose the treatment most appropriate for a given patient; (b) giving a patient, student, or customer a realistic estimate of how much of an effect is expected and how much effects differ from person to person; and (c) selecting the applicant who is most likely to best perform a specialized job. These investigations can also have pragmatic applications at the population level, for example (a) identifying populations that are most likely to benefit from psychological interventions and programs; and (b) choosing which interventions or programs are best suited to the subpopulations of interest. In summary, understanding how different people respond differently to experimental treatments and manipulations has profound implications for both basic scientific understanding and applied real-world problems.

Inferences about individual differences in causal effects, however, are complicated by the existence of uncontrolled extraneous variables, what Campbell and Stanley (1963) have referred to as *validity threats*. Although it is well understood that validity threats can bias inferences regarding the average effect of an experimental manipulation and methods to exclude such bias are well established, there is much less appreciation in psychology for how validity threats can bias inferences regarding individual differences in the effects of experimental manipulations, and there has not been much work on how to control for such bias. This article has two goals. The first goal is to discuss and illustrate how inferences regarding individual differences in the effects of experimental manipulations can be biased by threats to validity. The second goal is to introduce some structural equation modeling methods that exploit the power of randomized designs to control for many different forms of bias regarding both the mean effects of and individual differences in the effects of experimental manipulations. In the next section, I define the problem at hand and use the

This article was published Online First July 11, 2011.

Elliot M. Tucker-Drob, Department of Psychology and Population Research Center, University of Texas at Austin.

The Population Research Center at the University of Texas at Austin is supported by a center grant from the National Institute of Child Health and Human Development (R24 HD042849).

Correspondence concerning this article should be addressed to Elliot M. Tucker-Drob, Department of Psychology, University of Texas at Austin, 1 University Station A8000, Austin, TX 78712-0187. E-mail: tuckerdrob@psy.utexas.edu

prototypical within-subjects design to help to illustrate how extraneous variables can complicate inferences about individual differences in the effects of experimental manipulations. I then discuss how multiple group structural equation models can be fit to the data produced by a number of standard, as well as more novel, randomized designs to make strong inferences about individual differences in the effects of experimental manipulations. Finally, I illustrate the strengths and weaknesses of the designs previously discussed with a Monte Carlo simulation study and provide some general conclusions.

Individual Causal Effects and Validity Threats to Causal Inference

Experiments are conducted to infer causal effects. An *individual causal effect* for a given participant can be conceptually defined as the difference between the outcome that would be observed if the participant were to be assigned to the manipulation (i.e., treatment) condition and the outcome that would be observed if that same participant were to be assigned to the comparison (i.e., control) condition (Holland, 1986; Rubin, 2005). For example, a researcher might have a hypothesis concerning the effect of a stimulant medication on cognitive functioning. Using the definition just given, this researcher could conceptualize the causal effect of the stimulant medication for a given individual as the difference between how that person would perform on a given reasoning test at a given point in time if he or she were to take the medication minus how that same person would perform on the same reasoning test at the same point in time if he or she were to instead take a placebo (e.g., a sugar pill). A positive value of this difference (i.e., medication performance minus placebo performance is greater than 0) would be consistent with a cognitive enhancement effect of the medication (Greely et al., 2008).

Ideally, researchers would like to be able to directly compute each individual's causal effect, such that they can calculate the average cognitive enhancement effect of the medication relative to the placebo, calculate the standard deviation of the individual cognitive enhancement effects in the sample (how much person-

to-person variation there is in the effectiveness of the medication), and calculate correlations between observed participant characteristics and the magnitude of the cognitive enhancement effect (identifying the people for whom the medication is most effective). However, the logistical constraints of reality dictate that both *potential outcomes* (i.e., performance in the medication condition and performance in the placebo condition) cannot be directly observed for a given individual at the same point in time and under equal levels of naïveté to measurement or to treatment (Holland, 1986; Rubin, 2005). A given person's individual causal effect, therefore, can never be directly computed. Holland (1986) has termed this the *fundamental problem of causal inference*. The conditions could of course be administered to the same participant sequentially, but this approach has the potential to introduce a great deal of ambiguity to the situation.

To illustrate how causal inference becomes ambiguous when the same individuals are measured under both conditions, consider a prototypical within-subjects design, in which all participants are first measured in the comparison condition and are then measured in the manipulation condition. This design is schematized in the top portion of Table 1. To return to the medication example, a researcher using a within-subjects design might administer a reasoning test to the same group of participants on 2 consecutive days. One hour before taking the reasoning test on Day 1, each participant would take a sugar pill. One hour before taking the reasoning test on Day 2, each participant would take a pill containing the medication. To estimate the individual causal effect for each participant, the experimenter would simply calculate participant-specific difference scores (medication performance minus sugar pill performance). Positive values would be consistent with a cognitive enhancement effect for a given individual. The mean of these difference scores might be used as an index of the causal effect of the medication (relative to the placebo) on cognitive performance for the average or typical individual. Additionally, the standard deviation, or variance, of these difference scores might be used as an index of how much person-to-person variation exists in the magnitude of this causal effect. Finally, person-specific correlates (e.g., age) of the difference scores might be used

Table 1
Standard Designs

Group	First measurement	Second measurement
Simple within subjects		
1	Comparison	Manipulation
Simple between subjects		
1	Comparison	
2	Manipulation	
Between × Within		
1	Comparison	Comparison
2	Comparison	Manipulation
Counterbalanced position		
1	Comparison	Manipulation
2	Manipulation	Comparison
Counterbalanced forms		
1	Comparison (A)	Manipulation (B)
2	Comparison (B)	Manipulation (A)

Note. A and B are test forms.

to make inferences about whose cognitive performance benefits more or less than others from the medication (for formal treatments of moderation in within-subject designs of this sort, see Judd, Kenny, & McClelland, 2001; Judd, McClelland, & Smith, 1996). Although such a within-subjects approach is conceptually straightforward, it is unfortunately wrought with ambiguity, so much so that in one of their seminal articles on research design, Campbell and Stanley (1963) provided it as a “‘bad example’ to illustrate several of the confounded extraneous variables” (p. 7) that can bias causal inference. Campbell and Stanley, as well as more recent methodologists (e.g., Shadish, Cook, & Campbell, 2002), have primarily focused on how extraneous variables (i.e., internal validity threats) can bias estimates of average causal effects, and there do not appear to be any comprehensive discussions on how validity threats can bias inferences regarding individual differences in causal effects. I therefore provide such a discussion here.¹

The first problematic aspect of the within-subject design is that the outcomes associated with the comparison and the manipulation conditions are measured at different points in time. This introduces the possibility that other influences, apart from the causal effect, may be manifest in each individual’s difference score. Extraneous influences on the outcome that occur over time and are external to the individual are termed *history threats*. History threats include specific events (e.g., a natural disaster, the birth of a child, the weather, an e-mail from a friend) that occur concomitantly with the experimental manipulation that might affect the measured outcome. History can bias the estimate of the average causal effect if the events systematically affect all individuals over the course of the experiment. For instance, if Day 1 is a clear sunny day, and Day 2 is a dark rainy day, the average cognitive performance in the manipulation condition on Day 2 might be attenuated (perhaps because dreary days reduce participant motivation), leading to attenuation of the estimate of the average cognitive enhancement effect. History can also bias the estimated magnitude of individual differences in (i.e., the variance of) the causal effect if different events occur for different individuals or if individuals are differentially affected by the same event or events. For instance, individual differences in how much sleep the participants get between Day 1 and Day 2 might result in added variation in Day 2 performance (and, hence, in the Day 2 minus Day 1 difference score) that is not associated with variation in the individual cognitive enhancement effect of the medication. Finally, history can bias the estimated correlation between the causal effect and other variables. For example, if older children get less sleep than younger children between Day 1 and Day 2 (perhaps because of a late-night TV show that is popular among adolescents), age might be associated with lower difference scores, leading the researcher to incorrectly infer that the medication is less effective for older children.

Extraneous influences on the outcome that occur over time and are internal to the individual are referred to as *maturation threats*. Maturation includes processes such as hunger, fatigue, and psychological development. To the extent that a systematic maturational influence affects all people, estimates of the average causal effect will be biased. To the extent that individuals differ from one another in maturation, the estimated variance of the causal effect will be biased. Finally, to the extent that individual differences in maturation correlate with measured variables, the estimated cor-

relations among those measured variables and the individual causal effects will be biased. In our hypothetical example, if Day 1 is a Tuesday and Day 2 is a Wednesday and if individuals tend to become fatigued over the course of the week (thereby affecting their test performance), the estimate of the average cognitive enhancement effect of the medication might be downwardly biased. If different people become fatigued to different extents, the estimated variance of the cognitive enhancement effect could become inflated. Finally, if older children tend to experience this fatigue more than younger children, age might be associated with lower difference scores, leading the researcher to incorrectly infer that the medication is less effective for older children.

The second problematic aspect of the within-subject design stems from the fact that participants experience two conditions and are measured twice. When the participant is measured for the second time, he or she is not as naïve to the experiment or to being measured as he or she was when initially measured. Going back to the example, on Day 1 when participants take the placebo and then perform the cognitive task, they have never had any experience with the experiment, but on Day 2, when participants take the medication and then perform the cognitive task, they have already performed the task once before and have already had the experience of taking the placebo. Any effects that the experiences from Day 1 might have on performance on Day 2 are referred to as *reactivity*. Reactivity includes practice effects from having been exposed to the same measurement instrument previously, transfer effects to alternate measurement instruments, or any differences in behavior that may result from the participant figuring out the study or becoming sensitized to certain aspects of the tasks. For example, in our hypothetical experiment, participants might improve on the cognitive task from the first to the second assessment simply because they are familiar with it, thereby potentially distorting the value of the mean difference score. If some people benefit more than others from having been previously tested, then the variance of the difference scores and the observed pattern of correlates between the difference scores and other variables may not exclusively reflect individual differences in, and predictors of, medication-related cognitive enhancement but rather may partially reflect individual differences in and predictors of the effects of retest-related learning (e.g., Salthouse & Tucker-Drob, 2008). It is possible that changing the cognitive measure from the first day to the second day may help to reduce participant familiarity with the test and hence may help to reduce reactive effects. However, this can introduce an *instrumentation threat*, in that the different measurement instruments may lack comparability (measurement equiva-

¹ Validity threats that are not discussed in this article include *selection*, *measurement*, and *mortality/attrition*. Selection, in which pre-existing differences in means, variances, and covariances are associated with the nonrandom assignment of participants to groups, is not an issue for the single-group within-subjects design and the multiple-group randomized designs that are the focus of this article. Measurement, which refers to differential difficulty or sensitivity of a given measurement instrument across individuals or testing occasions, is not directly relevant to this article, in that it is a property of the instrument rather than a specific design. Finally, nonrandom dropout of participants that is due to selective mortality or attrition are potential threats to internal validity for all designs in which participants are measured more than once.

lence).² Finally, it is important to note that although reactive effects include any differences in performance that result from having been previously measured or from having experienced aspects of the experiment previously, they differ conceptually from *carryover effects*, which refer to genuine causal effects of the manipulation that persist across measurement occasions. An example of a carryover effect is the possibility that taking a medication on Day 1 has a lasting cognitive enhancement effect that is still evident (although perhaps not as strong) on Day 2, when the participant is measured for a second time.

Multiple measurements of each individual also serve to compound imprecision of measurement. Psychometric theory virtually guarantees that the measured outcome will not be a perfect reflection of the trait of interest but will also contain transient and unsystematic influences (e.g., measurement error) that differ from person to person and vary randomly from measurement to measurement and occasion to occasion. Any difference score that is calculated between two observed outcomes will inevitably contain between-person variation in these unsystematic influences (Cronbach & Furby, 1970), which, for the simple within-subjects design, will serve to inflate the estimated variance of the causal effect. Moreover, because these influences cause some people to score more extremely and others to score less extremely than their true (time invariant) scores on the construct of interest, they can produce a more negative (or less positive) relation between initial scores and change, which is termed *regression to the mean*.³ For example, suppose that a given person has a true score of seven on the reasoning test when taking the sugar pill and has a true score of nine on the reasoning test when taking the medication. Further, suppose that this person makes a lucky guess on the reasoning test on Day 1 and therefore scores an eight in the sugar pill condition. This person is not very likely to make another lucky guess on Day 2 and might therefore score a nine in the medication condition. This person's difference score would be 1, even though the causal effect is truly a two. One can further imagine another person who got unlucky and scored less than his or her true sugar pill condition score of seven on Day 1 and then scored closer to his or her true medication condition score of nine on Day 2. This person's difference score would be higher than the true causal effect. The net result would be (a) a downwardly biased estimate of the relation between comparison performance and the magnitude of the causal effect and (b) an upwardly biased estimate of the variance of the causal effect.

Ensuring Internal Validity Through Randomization

The randomized experiment is social science's most revered approach to producing accurate estimates of causal effects of experimental manipulations (Campbell & Stanley, 1963; Fisher, 1925; McCall, 1923; Rubin, 2005; Shadish, Cook, & Campbell, 2002). Randomizing participants to groups that experience different conditions ensures that, within the bounds of sampling fluctuation, individual differences (in both traits and exogenous experiences) are evenly distributed across the groups, such that any observed differences between the groups can be attributable to differences in the conditions. For the simple between-subjects design, in which participants are randomly assigned to a single measurement under either the comparison condition or the manipulation condition (see Table 1), the standard implication is that,

under very few and often highly plausible assumptions (e.g., that participants do not influence one another; Rubin, 2005), the difference between the average outcome in the manipulation condition and the average outcome in the comparison condition will be an unbiased estimate of the average of the individual causal effects in the population.

Perhaps because causal inference in randomized experiments is based on the premise that individual differences and idiosyncrasies average out across groups, conventional experimental methodology predominantly focuses on estimating population-average causal effects and has largely neglected questions concerning person-to-person variation in the magnitudes of individual causal effects and their correlates. However, although not widely recognized, just as randomization ensures that, *ceteris paribus*, group means will be equal under the null hypothesis, it also ensures that within-group variances, covariances, and regression relations will be equal under the null hypothesis. In this section, I demonstrate how one can begin to build statistical models that capitalize on these added properties of randomization, such that variance and covariance components of the causal effect can be confidently estimated.

Simple Between-Subjects Design

This design is the most basic randomized experimental design. As described earlier and schematized in Table 1, this design involves the random assignment of participants to one of two groups, with one group experiencing the manipulation condition and the other group experiencing the comparison condition. For our medication example, this would entail randomly assigning participants to either a group that takes a sugar pill and is then administered the reasoning test or a group that takes the medication and is then administered the reasoning test. The meticulous researcher would ensure that all participants took the same reasoning test at the same time under the same conditions, perhaps by administering the test to all participants in the same room after randomly handing out unmarked pills to them after they were seated. The first thing to note about this design is that, by not measuring any given participant more than once, many of the validity threats described earlier are entirely avoided. That is, because the different conditions are not separated by time, history and maturation threats do not factor in, and because participants are not measured twice, regression to the mean and reactivity are not issues. However, because participants are not measured under both conditions, individual causal effects (i.e., manipulation–comparison difference scores) cannot be directly computed. As such, causal inference must be made through across-person comparisons.

² Differences in the difficulties of the measurement instruments (i.e., intercepts or response thresholds) can potentially bias mean effects, whereas differences in the sensitivities of the measurement instruments (i.e., discrimination, communality, or reliability) can potentially bias individual differences.

³ Both systematic and unsystematic sources of time-specific variance can result in regression to the mean. One such source is systematic within-person occasion-to-occasion fluctuation, also known as intraindividual variability (see, e.g., DeShon, 1998; Salthouse, 2007).

Typically, researchers using the simple between-subjects design are primarily concerned with testing for an overall average causal effect, which they do using the t test (or analysis of variance [ANOVA] for more complex designs that include multiple manipulations). Cohen (1968) has shown how a t test can be parameterized as a linear regression, written here as follows:

$$Y = b_0 + b_1 \times g + u, \quad (1)$$

where Y is the measured outcome, g is a dummy coded variable representing group membership (comparison and manipulation conditions are coded as 0 and 1, respectively), the regression intercept (b_0) is equal to the mean level of performance in the comparison condition, and the regression coefficient (b_1) is equal to the mean difference in performance between manipulation and comparison conditions. With this approach, individual differences in the magnitude of the causal effect cannot be directly estimated. However, although such an approach is not very commonly implemented, it is rather straightforward for researchers using this design to test whether individual causal effects relate to measured participant characteristics. This approach, which was pioneered by Cronbach (see, e.g., Cronbach, 1975), involves testing whether the regression slope relating the measured outcome (Y) to group membership (g), differs according to a person's score on a measured characteristic, x . This can be achieved by including terms for the main effects of x and the interaction of x with the grouping variable, g , in the regression predicting the outcome, Y .

$$Y = b_0 + b_1 \times g + b_2 \times x + b_3 \times x \times g + u. \quad (2)$$

If the interaction term, b_3 , is statistically significant, this would be evidence for what Cronbach termed an *Aptitude \times Treatment interaction*, where aptitude is defined as "any characteristic of the person that affects his response to the treatment" (Cronbach, 1975, p. 116). To make this more concrete, if x were age, g represented medication versus sugar pill, and Y represented reasoning performance, the b_3 parameter would reflect the extent to which the cognitive enhancement effect differed linearly with age. This approach is very similar to including grouping or blocking variables (e.g., gender or age group) as variables in an ANOVA context (see, e.g., Kirk, 1995). Both the regression and the ANOVA approaches to examining measured participant characteristics as correlates of (i.e., moderators of) causal effects produce estimates of what might be termed *conditional* (or *marginal*) average causal effects, for example, the average causal effect for women or the average causal effect for 11-year-old children. That is, they effectively produce average causal effects for population subgroups (Rubin, 2005; Steyer, Nachtigall, Wüthrich-Martone, & Kraus, 2002).

One outstanding question is whether random individual differences in the causal effect (i.e., individual differences that may not be accounted for by measured covariates) can be estimated from the data produced from the simple between-subjects design and, if so, under what assumptions. Estimating the variances of random or latent variables representing causal effects is important for a number of related reasons. First, if observed variables do not appreciably modify the size of the causal effect, individual differences in the causal effect may still be large but simply difficult to predict. For both applied and theoretical reasons, it may be important to know how much heterogeneity there is, even if this heterogeneity

cannot be accounted for (e.g., How certain can a doctor be about the magnitude of an effect to expect when prescribing a pill to a patient? To what extent are a basic scientist's new findings indicative of a nomothetic principle that governs how all humans behave?). Second, it may be useful to examine what proportion of individual differences in the causal effect is accounted for by observed variables, and to do so requires knowing what the total variance of the causal effect is. Third, identifying causal effects on multiple outcomes as random coefficients or latent variables is necessary to examine whether they correlate with one another. Finally, there is an accumulating literature demonstrating that the existence of individual differences in causal effects can serve to undermine standard approaches to examining causal mediation (Bauer, Preacher, & Gil, 2006; Glynn, 2010; Kenny, Korchmaros, & Bolger, 2003). Estimating individual differences in causal effects can therefore be used to test an important assumption of causal mediation and perhaps even relax it.

With group equivalence of variances of the outcome as a null hypothesis, one can examine whether the manipulation and comparison groups differ in the magnitudes of their variances (Bryk & Raudenbush, 1988). Going back to the example, one might find that concomitant with mean advantages in reasoning performance for the medication group relative to the sugar pill group, the medication group is also more heterogeneous in reasoning performance than the sugar pill group is (i.e., the variance in reasoning performance is larger for the medication group than it is for the sugar group). This would be evidence for individual differences in the causal effect. However, the between-group difference in (residual) variances will be an unbiased estimate of the variance of the causal effect only if the causal effect is statistically independent of scores in the control condition (see Appendix A for a proof), conditional on any measured covariates. Because participants are not exposed to both manipulation and comparison conditions, this covariance cannot be estimated. To make this concept more concrete, cognitive performance in the sugar pill condition might be correlated with the cognitive enhancement effect of the medication. Not only can this correlation not be estimated from data produced by a simple between-subjects design, but if it is truly positive, the across-group difference in variance will be an overestimate of the true variance of the cognitive enhancement effect of the medication (the researcher will conclude that the cognitive enhancement effect of the medication differs from person to person to a larger extent than it truly does). Researchers using the simple between-subjects design must therefore appraise the tenability of the assumption that the causal effect is statistically independent of performance in the control condition on theoretical grounds when deciding whether the variance subtraction method is trustworthy.

An integration of these concepts serves as the basis for the first structural equation model introduced in this article. This structural equation model is depicted as a path diagram in Figure 1. This figure has a number of features that are used in many of the path diagrams presented in this article. Measured variables are depicted as squares, with Y representing the experimental outcome (e.g., reasoning performance) and x representing a measured participant characteristic (e.g., age). Latent variables are represented as circles, with F_c representing performance in the comparison condition and F_Δ representing the individual causal effect (the subscript Δ was intentionally cho-

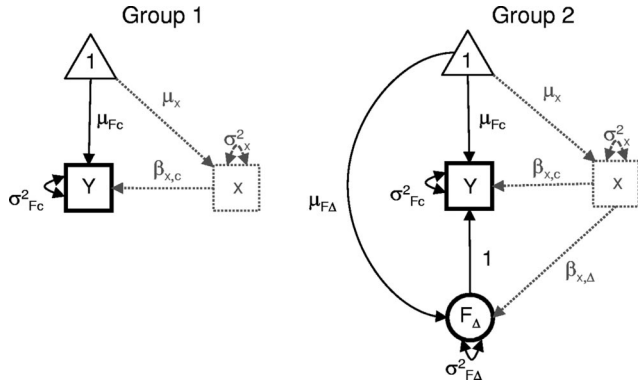


Figure 1. Structural equation model for the simple between-subjects design. See Table 1 for a schematization of how data are collected for this design, Table 2 for a glossary of symbols used, and the in-text description for further details.

sen so that attention is drawn to the fact that the causal effect is conceptualized as a within-person difference between comparison and manipulation performance). Unidirectional arrows represent regression relationships, and bidirectional arrows represent variances or covariance relationships. Numbered paths

(arrows) are fixed to the values shown, and labeled paths are estimated from the data, with paths having the same label constrained to be equal to one another. The triangle represents the unit constant that is used to estimate variable means and intercepts. Note that inclusion of a covariate is not necessary in any of the structural equation models discussed in this article. Therefore, in all figures, the terms involving the covariate are represented in light grey dotted lines. To aid readers with the interpretation of the path diagrams displayed in this article, a glossary of symbols is presented in Table 2. This glossary provides a psychological description of the meaning of each symbol.

The key features of the structural equation model in Figure 1 are as follows: First, the observed mean and variance of the outcome for participants assigned to the comparison condition reflect the mean (μ_{F_c}) and variance ($\sigma_{F_c}^2$) of the theoretical comparison condition performance. Second, the observed mean of the outcome for participants assigned to the manipulation condition is equal to the mean of the causal effect (μ_{F_Δ}) plus the mean comparison condition performance (μ_{F_c}). Third, the observed variance of the outcome for participants assigned to the manipulation condition is equal to the variance of comparison condition performance ($\sigma_{F_c}^2$) plus the variance of the causal effect ($\sigma_{F_\Delta}^2$). Fourth, the magnitude of the regression of the

Table 2
Glossary of Symbols Used in Path Diagrams

Symbol	Description
Variables	
Y	Observed outcome (e.g., reasoning test performance)
F_c	Inferred true score in the comparison condition (e.g., the score that the participant would receive if he or she took the reasoning test in the sugar pill condition, naïve to previous measurement or treatment)
F_Δ	Inferred causal effect = theoretical true score in manipulation condition – theoretical true score in the comparison condition (e.g., reasoning performance in medication condition – reasoning performance in sugar pill condition)
F_T	Inferred net effect of extraneous variables (e.g., history, maturation, reactivity, measurement error). T stands for threat to internal validity.
x	A measured covariate (e.g., age)
Parameters	
μ_{F_c}	Mean of the inferred true score in the comparison condition
μ_{F_Δ}	Mean of the inferred causal effect
μ_{F_T}	Mean net effect of extraneous variables
$\sigma_{F_c}^2$	Between-person variance of the inferred true score in the comparison condition
$\sigma_{F_\Delta}^2$	Between-person variance of the inferred causal effect
$\sigma_{F_T}^2$	Between-person variance of the net effect of extraneous variables
$\beta_{c,\Delta}$ or $\sigma_{c,\Delta}$	Regression or covariance between individual differences in true comparison condition performance and individual differences in the causal effect
$\beta_{c,T}$ or $\sigma_{c,T}$	Regression or covariance between individual differences in true comparison condition performance and individual differences in the net effect of extraneous variables
$\beta_{\Delta,T}$ or $\sigma_{\Delta,T}$	Regression or covariance between individual differences in the causal effect and individual differences in the net effect of extraneous variables
$\beta_{x,c}$ or $\sigma_{x,c}$	Regression or covariance between a measured covariate and individual differences in true comparison condition performance
$\beta_{x,\Delta}$ or $\sigma_{x,\Delta}$	Regression or covariance between a measured covariate and individual differences in the causal effect
$\beta_{x,T}$ or $\sigma_{x,T}$	Regression or covariance between a measured covariate and individual differences in the net effect of extraneous variables
λ_w	Factor loading of test form w on true performance
v_w	Intercept of test form w
σ_w^2	Residual variance of test form w
α	Carryover of the causal effect from having been exposed to the manipulation condition at a previous occasion

outcome in the manipulation group on participant characteristic x is equal to the magnitude of the regression of the outcome in the comparison group on x ($\beta_{x,c}$) plus the magnitude of the regression of the causal effect on x ($\beta_{x,\Delta}$; this term is equivalent to an $x \times \text{Group}$ interaction term, i.e., $\beta_{x,\Delta}$ is directly analogous to the b_3 coefficient in Equation 1). Fifth, conditional on the covariate, x , performance in the comparison condition and the individual causal effect are uncorrelated. In many cases, the fifth assumption may not be considered tenable. When the actual comparison performance–causal effect correlation is not zero but is modeled as such for the purposes of model identification, the structural equation model depicted in Figure 1 will produce a biased estimate of the variance of the causal effect (see Appendix A). Moreover, estimating the comparison performance–causal effect correlation may, in fact, be of substantive interest to the experimenter (e.g., in determining whether unmedicated individual differences in reasoning ability are related to the cognitive enhancement effects of the medication). A design that allows for the covariance between the outcome in the control condition and the causal effect to be estimated is therefore discussed next.

Between \times Within Design

This design (also sometimes referred to as the *randomized pretest–posttest design*) combines many of the advantages of the simple between-subjects design with those of the simple within-subjects design. In this design, participants are randomly assigned to one of two groups, each of which is measured on two occasions (see Table 1). Participants in Group 1 experience the comparison condition twice, whereas participants in Group 2 first experience the comparison condition and then experience the manipulation condition. Notice that the participants in Group 2 experience both conditions, just in the simple within-subject design. As discussed earlier, this has the advantage of allowing for both the comparison and the manipulation outcomes to be observed on the same individuals, but if Group 2 were the only condition, this would also have the disadvantage of introducing a number of extraneous influences (validity threats) associated with the passage of time (history and maturation) and with repeated measurements (reactivity and regression to the mean). In the Between \times Within design, Group 1 serves as a control for these extraneous influences. That is, all of the influences associated with the passage of time and repeated measurements (i.e., history, maturation, reactivity, and regression to the mean) are reflected in the changes observed in Group 1, whereas all of these influences and the effects of the manipulation are reflected in the changes observed in Group 2. As such, any between-group differences in means, variances, or covariance/regression relations that are observed in the patterns of Occasion 1 to Occasion 2 changes can be associated with the causal effect.

An example of a Between \times Within design might entail randomly assigning participants to either (a) a group that takes a sugar pill and a reasoning test on Day 1 and then repeats this process on Day 2 or (b) a group that takes a sugar pill and a reasoning test on Day 1 and then takes the medication and the same reasoning test on Day 2. If the correlation between Day 1 performance and the Day 2 minus Day 1 difference score differs across groups, this would be evidence that performance in the comparison condition (sugar pill condition) is truly correlated with the causal effect. This

can be tested with a multiple regression model in which Y_2 (Day 2 performance) is predicted by Y_1 (Day 1 performance), g (a dummy coded grouping variable in which Group 1 = 0 and Group 2 = 1), and the interaction of Y_1 with g :

$$Y_2 = b_0 + b_1 \times g + b_2 \times Y_1 + b_3 \times Y_1 \times g + b_4 \times x + b_5 \times x \times g + u, \quad (3)$$

with the test of b_3 being analogous to a test of heterogeneity of regression in an analysis of covariance. In Equation 3, a b_3 coefficient that is significantly different from zero would indicate that performance in the comparison condition (sugar pill condition) is correlated with the causal effect. Note that the terms in parentheses in the Equation 3 can be included to test whether a measured covariate, x , relates to the causal effect, just as was discussed for the simple between-subjects design. A similar formulation of the Equation 3 regression explicitly models the Y_2 minus Y_1 difference as the outcome of interest:

$$\Delta Y = Y_2 - Y_1 = b_0 + b_1 \times g + b_2 \times Y_1 + b_3 \times Y_1 \times g + b_4 \times x + b_5 \times x \times g + u. \quad (4)$$

It is important to keep in mind that the Y_2 minus Y_1 difference in Group 1 reflects change due to extraneous influences, whereas the corresponding difference in Group 2 reflects both these extraneous influences plus the individual causal effect. As such, the between-group difference in the mean difference score is an unbiased estimate of the average causal effect, the between-group difference in the regression of the difference score on Y_1 (i.e., the b_3 interaction term) is an unbiased estimate of the regression of the causal effect on comparison condition performance, and the between-group difference in the regression of the difference score on a measured covariate (i.e., the b_5 interaction term) is an unbiased estimate of the regression of the causal effect on the covariate. Moreover, the between-group difference in the (residual) variance of the difference score is an unbiased estimate of the (residual) variance of the causal effect, assuming that the causal effect is uncorrelated with the extraneous influences (see Appendix B) conditional on the covariates. To illustrate, in our example, the Group 2 minus Group 1 difference in the variances of the Day 2 minus Day 1 difference score is an unbiased estimate of the variance of the cognitive enhancement effect of the medication, assuming that the magnitude of the cognitive enhancement effect is uncorrelated with individual differences in history, maturation, and reactive effects. In many cases this is an acceptable assumption. For example, our hypothetical researcher may find it unlikely that the extent to which participants benefit from the experience of having taken the reasoning test before (e.g., the retest effect) correlates with the cognitive enhancement effect that they get from the medication.

An integration of these concepts serves as the basis for the structural equation model depicted in Figure 2. In both groups, Y_2 is regressed onto Y_1 at a fixed value of 1, such that all remaining predictors of Y_2 can be interpreted as predictors of the Y_2 minus Y_1 difference score (McArdle & Nesselrode, 1994). This model (cf. Sörbom, 1978; Steyer, 2005) is quite unique for analyzing experimental data in that, in addition to including a factor representative of the causal effect of the manipulation, it also explicitly includes

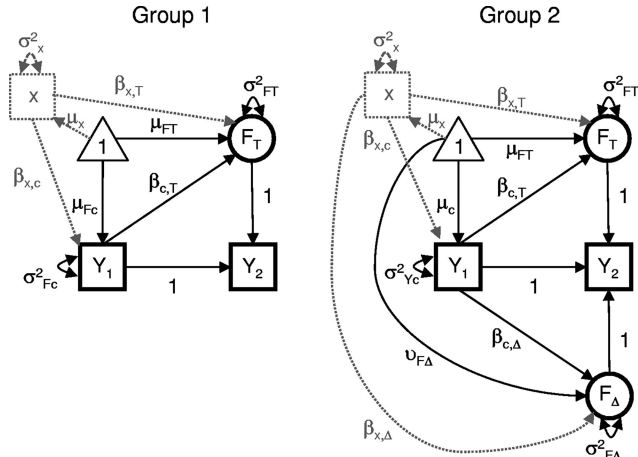


Figure 2. Structural equation model for the Between \times Within design. The subscripts on Y correspond to the first measurement (1) and the second measurement (2). See Table 1 for a schematization of how data are collected for this design, Table 2 for a glossary of symbols used, and the in-text description for further details.

a factor (F_T) representative of extraneous influences (i.e., validity threats).⁴ This factor is identified from the across-occasion change in performance (i.e., the Y_2 minus Y_1 difference score) that is observed in Group 1. As mentioned earlier, because this group experiences the comparison condition twice and never experiences the manipulation condition, any patterns of change in performance that are observed from Occasion 1 to Occasion 2 in this group necessarily constitute changes that are extraneous to the experimental manipulation. Alternatively, because Group 2 experiences the comparison condition on the first occasion and the manipulation condition on the second occasion, the patterns of change in performance that are observed from Occasion 1 to Occasion 2 are composed of the changes attributable to the actual causal effect of the manipulation in addition to changes that are extraneous to the manipulation. Both groups therefore contain an F_T factor representative of extraneous influences, or validity threats. The mean, variance, and regression/covariance relationships related to F_T are constrained to be equivalent in both groups. This is an assumption that is ensured by the fact that the threat-related experiences (e.g., time elapsing, participants being tested twice) in both groups can be assumed to be equivalent combined with the fact that participants are randomized to the two groups (such that individual differences are evenly distributed across groups). This is indicated in Figure 2 by the fact that corresponding terms are given same label in both groups. Only the second group, however, contains an F_{Δ} factor, representative of the causal effect of the manipulation. This factor accounts for the mean Y_2 minus Y_1 difference, the variance of this difference, and the covariance/regression relationships involving this difference that are not accounted for by F_T . In other words, F_{Δ} accounts for the patterns that differ between groups. Note that, for identification purposes, the covariance between extraneous factors (F_T) and the causal effect (F_{Δ}) cannot be estimated with this method. As is analytically demonstrated in Appendix B and empirically illustrated in the simulation study reported later, this can potentially lead to biased estimates of the variance of the causal effect ($\sigma_{F_{\Delta}}^2$). To make this concrete, this

model is unable to determine whether those participants who receive the largest cognitive enhancement effect from the medication tend to be the same participants whose scores benefit the most from the previous experience of taking the reasoning test. If such a positive correlation exists, the estimated variance of the cognitive enhancement effect will be upwardly biased.

Test Equating for Experiments

In many research areas, the dominant threat to internal validity is participants' reactivity to being retested on the same material. At the same time, the measurement of individuals more than once produces important information about changes that occur within individuals as they proceed through varying aspects of the experiment. One possible way to produce within-person estimates of within-person differences while avoiding threats associated with reactivity to retesting might be to use different measurement materials for each phase of measurement. The main problem with such a research approach, however, is that it results in outcomes that are not easily comparable (an *instrumentation threat*). In this section, test-equating approaches are reviewed, and new methods to integrate them into the experimental paradigm to produce comparable nonrepeated measures are discussed. In a later section, test-equating procedures are integrated with designs that allow for the effects of history, maturation, and any remaining reactive effects to be separated from the causal effect.

Test-equating procedures stem from a perspective that is foundational to both classical and modern psychometric theory: Observed levels of performance on a given measure are imperfect indications of unobserved (latent) traits that can be measured at least as well with many alternative materials and/or methods. By constructing data-based models of the relations between the unobserved (assumed) trait of interest and observed scores, researchers can establish a more valid and generalizable network of relational patterns between the trait and its correlates and, as a byproduct, can produce inferences about the common trait, using a number of alternative materials and methods of measurement. This byproduct can be used advantageously to measure individuals on the same outcome multiple times without ever repeating the actual method of measurement. Reactive effects, therefore, can be potentially reduced without producing the instrumentation threats that normally would be associated with using different measures at different phases of the experiment.

Data collection designs for three basic forms of test equating are schematized in the top portion of Table 3.⁵ These can be characterized as common person equating, common test equating, and equating purely by randomization (Angoff, 1971; Crocker & Algina, 1986; Kolen & Brennan, 2004; Masters, 1985). Common

⁴ See B. O. Muthén and Curran (1997) for a similar approach, in which treatment effects are distinguished from normative developmental trajectories.

⁵ Note that for two reasons, Table 3 does not specify the sequence in which the tests are administered. First, test equating is introduced here as a means of reducing the effects of the sequences of measurement. Second, this section introduces the basic elements of test equating so that they can, in a later section, be incorporated into a more general framework that does take sequences of measurement into account.

Table 3
Designs for Test Equating and Test Equating for Experiments

Group	Test A	Test B	Test D
Common person equating			
1	X		
2		X	
3	X	X	
Common test equating			
1	X		X
2		X	X
Equating by randomization			
1	X		
2		X	
Common test equating for experiments			
1	Comparison	Manipulation	Comparison
2	Manipulation	Comparison	Comparison

Note. An X denotes a measurement.

person equating involves calibrating two (or more) tests to the same group of people, such that when administered in the absence of one another, the tests produce scores that are on the same metric. Common test (or common item) equating involves administration of an anchor test (Test D) to each group, in addition to that group’s unique test. The group-specific tests are then calibrated relative to the anchor test, such that all scores are again on the same scale of measurement. Equating purely by randomization involves administration of separate tests to groups that have been randomly assigned. Because it can be assumed that the randomization has produced groups that do not differ in the mean and distribution of their true scores on the construct measured by the two tests, the test scores can each be converted to a common metric (e.g., the standardized *z* score metric).

Common Test Equating for Experiments

This novel design, which is schematized in the bottom portion of Table 3, is characterized by (a) each participant being tested no more than once on Test Forms A, B, and D; (b) one group in which Test Form A is paired with the manipulation condition and Test Form B serves as the comparison measurement; (c) one group in which Test Form B is paired with the manipulation condition and Test Form A serves as the comparison measurement; and (d) in both groups, Test Form D serving as a common anchor to which Test Forms A and B can be calibrated. The anchor test (Test Form D) allows the experimental outcomes to occur on a common metric, such that the causal effect can be deconfounded from instrumentation artifacts associated with differential sensitivities and/or differential difficulties of the different measurement materials. To make this idea concrete, an experiment involving cognitive enhancement effects might entail randomly assigning participants to either (a) a group in which they took Reasoning Tests A and D after taking a sugar pill and Reasoning Test B after taking the stimulant medication or (b) a group in which they took Reasoning Tests B and D after taking a sugar pill and took Reasoning Test A after taking the stimulant medication. All participants always experience the sugar pill and the medication conditions, but no one takes the same reasoning test twice. Not repeating the same

measurements on a given individual may help to reduce any practice effects on the reasoning test that might otherwise confound the calculated medication–sugar pill difference score. Further, because all individuals are tested on Anchor Test D, their scores on Tests A and B can be calibrated to a common metric, such that meaningful (within-person across-condition) difference scores can be computed.

A structural equation model for the common test equating for experiments approach is displayed in Figure 3. In this figure, each test is represented by a single variable. In this model, the specific magnitudes of each test’s loading and intercept are allowed to differ according to the specific test form (A, B, or D) but are constrained to be invariant across groups. This is indicated in Figure 3 by all loadings (λ) and intercepts (ν) having subscripts that are specific to the test form. In both groups, all variables load on F_c , the factor representative of comparison condition performance. However, whether a given variable loads on the factor representing the causal effect (F_Δ) differs between groups depending on the condition that was paired with the test for the group. As such, Test B loads on the causal effect in Group 1 but not Group

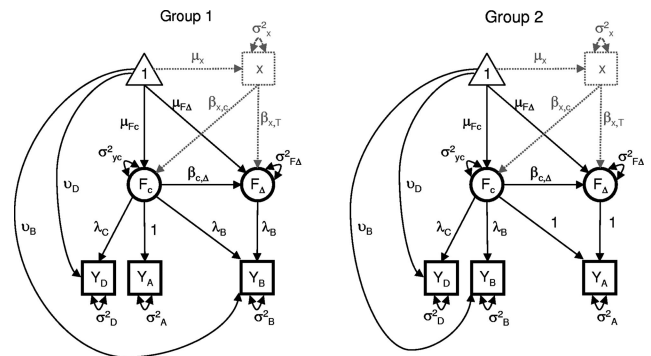


Figure 3. Structural equation models for common test equating for experiments. The subscripts on *Y* correspond to test forms A, B, or D. See Table 3 for a schematization of how data are collected for this design, Table 2 for a glossary of symbols used, and the in-text description for further details.

2, and Test A loads on the causal effect in Group 2 but not Group 1. This amounts to a simple within-subjects design specified to occur at the factor level (i.e., the manipulation–comparison difference score is calculated from factors rather than manifest observations). As in the simple within-subjects design, this approach does not include provisions for the estimation of factors representative of time- or sequence-related changes. However, compared with the simple within-subjects design, this approach has the advantage of never repeating measurements of the same participant with the same test, thereby potentially reducing reactive effects. Later on, a more complex design is introduced that combines the advantages of test equating for nonrepeated measurements with those of the Between \times Within design for controlling for time- and sequence-related changes.

A General Framework for Experiments

Procedures have been reviewed that demonstrate how one can begin to separate both the mean effects of and individual differences associated with the passage of time, the sequences of measurement, and the specific measurement materials used, from those associated with the actual causal effect of the experimental manipulation. Whereas the preceding statistical models have been in path diagram form for specific data collection methods, a general equation-based model is presented here to represent how all three influences (manipulation condition vs. comparison condition, sequence/time, and test form/measurement materials) operate in an experiment:

$$Y_{w,m,p,n} = v_w + \lambda_w \times F_{c,n} + m \times \lambda_w \times F_{\Delta,n} + p \times \lambda_w \times F_{T,n} + u_{w,p,m,n} \quad (5)$$

This model explains that the score on measure w for person n , administered in position p in the presence or absence of the manipulation (m), is a function of a test-specific intercept (v), a factor representing individual differences in comparison condition performance (F_c), a factor representing the causal effect of the manipulation (F_Δ), an extraneous variable factor (or *threat* factor) representing the effects of validity threats associated with the sequence/time of testing (F_T), and an assessment-specific unique (residual) factor (u).⁶ The parameter λ is a test-specific scaling coefficient (factor loading). On the right side of the equation, m and p act as (typically dummy coded) coefficients that denote whether the material was accompanied with (1) or without (0) a manipulation and whether the test was administered first (0) or subsequently (1) in the sequence, respectively. With the exception of the unique factors, the factors each have their own means (μ , the average effects) and variances (σ^2 , individual differences in the effects) and for many designs are allowed to have covariances with one another (σ). For some designs, the unique factors can be allowed to have their own variances (σ_w^2). Conventional factor identification constraints (e.g., fixing a single loading to 1 and a single intercept to 0) are necessary.

Equation 5 makes explicit the rather straightforward assumptions on which each of the preceding analytical models (i.e., the path diagrams depicted in Figures 1–3) were constructed. First, the causal effect (F_Δ) affects performance only on measurements that have been paired with the experimental manipulation. Second, the

threat factor that is associated with extraneous variables (F_T) does not affect performance on the first measurement occasion and always affects performance on the subsequent measurement occasion. Third, test difficulty (the test intercepts), the extent to which the tests reflect the latent outcome (the factor loadings), and errors of measurement (the variances of the unique factors) are properties of the test, rather than the person, such that they are invariant across the groups or conditions. It follows from these assumptions that the presence versus absence of the experimental manipulation, the sequential positions of measurement, and the measurement instruments, combine to produce individual levels of performance on the outcome of interest, Y . Note that, although not represented in Equation 5, the comparison condition performance (F_c), the causal effect (F_Δ), and the net effect of extraneous variables (F_T) can be regressed on (or allowed to covary with) other measured variables or latent factors for which data may be available.

The path diagrams displayed in this article can all be considered instantiations of Equation 5, with specification of the m and p coefficients to correspond to each specific design's features and with constraints placed on the F_c , F_Δ , F_T , and u factor variances and covariances to ensure model identification. Such design-specific parameter specifications and constraints can be found in Table C1 of Appendix C. Table C1 also contains Equation 5 specifications for the two advanced experimental designs that are discussed next. These designs integrate many of the advantageous features of the preceding designs (e.g., randomization, a comparison condition control group, multiple nonrepeated measurements), while allowing for identification of all components of the comparison condition performance (F_c), causal effect (F_Δ), and extraneous variable (F_T) factor variance–covariance matrix, thereby reducing potential estimation biases with respect to the causal effect.

Two Advanced Experimental Designs

The framework developed in this article enables the careful development of novel experimental designs that differ in their combinations of methods or materials of measurement, the presence versus absence of the key manipulation, and the sequences in which the measurements and manipulation presence versus absence occur. The specific design has direct implications for the parameters that can be estimated (identified) in the corresponding statistical model. Table 4 schematizes two designs that allow for identification of all three (F_c , F_Δ , and F_T) factors in Equation 5 and all covariances between them. As was the case for the standard experimental designs reviewed earlier, the following designs rely on randomized assignment of participants to conditions.

⁶ Some researchers may not be interested in analyzing individual causal effects per se but may rather be interested in analyzing individual differences in performance under two different experimental conditions. The current framework could be straightforwardly adapted for such purposes. Rather than modeling outcome Y as a function of a threat factor, comparison condition performance, and the causal effect of the manipulation/treatment, one would model Y as a function of a threat factor, Condition 1 performance, and Condition 2 performance.

Table 4
Two Advanced Experimental Designs

Group	First measurement	Second measurement
Three-group repeated measure		
1	Comparison	Comparison
2	Comparison	Manipulation
3	Manipulation	Comparison
Three-group nonrepeated measures		
1	Comparison (A)	Comparison (B)
2	Comparison (B)	Manipulation (A)
3	Manipulation (A)	Comparison (B)

Note. A and B are test forms.

Three-Group Repeated Measure Design

This design can be considered a further elaboration of the Between \times Within design. This design separates the effects of the key manipulation and having been previously measured by way of measurements in both the comparison and the manipulation condition and both with and without the experience of a previous measurement. Like the Between \times Within design, the three-group repeated measure design contains a comparison condition:comparison condition repeated measurement control group and a comparison condition:manipulation condition repeated measurement experimental group. The additional third group, a manipulation condition:comparison condition repeated measurement experimental group, helps to further deconfound the manipulation from the threat-related factor. This allows for identification of the correlation between the causal effect (F_{Δ}) and the net effect of extraneous influences (F_T) and can prevent a biased estimate of variance of the causal effect ($\sigma_{F_{\Delta}}^2$) that may arise in the Between \times Within design (see Appendix B). Application of this design to a cognitive enhancement experiment would entail randomly assigning participants to either (a)

a group that takes a sugar pill and a reasoning test on Day 1 and then repeats this process on Day 2, (b) a group that takes a sugar pill and a reasoning test on Day 1 and then takes the medication and the same reasoning test on Day 2, or (c) a group that takes the medication and a reasoning test on Day 1 and then takes a sugar pill and the same reasoning test on Day 2.

A three-group path-diagram representation of the application of Equation 5 to the data produced by the three-group repeated measure design is depicted in Figure 4. The subscripts on Y correspond to the first and second measurements. In parentheses underneath the Y variables are indications of whether the measurement was paired with the comparison condition (e.g., the sugar pill) or the manipulation condition (e.g., the medication). No manipulation condition is administered to Group 1; hence, the causal effect, F_{Δ} , does not affect performance on either Y_1 or Y_2 (this is equivalent to the m coefficient in Equation 5 taking on a value of 0 for both measurements). In Group 2, the causal effect influences the second measurement (Y_2) but not the first measurement (Y_1). Finally, in Group 3, the causal effect influences the first measurement (Y_1), and a carryover of this causal effect to the second measurement (Y_2) is freely estimated as α (i.e., the m

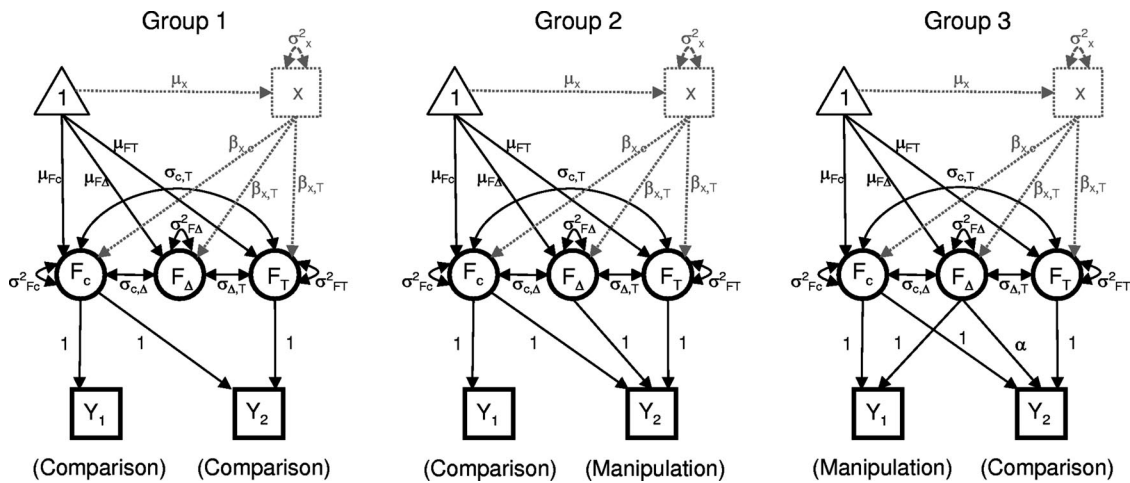


Figure 4. Structural equation model for the three-group repeated measure design. The subscripts on Y correspond to the first measurement (1) and the second measurement (2). See Table 4 for a schematization of how data are collected for this design, Table 2 for a glossary of symbols used, and the in-text description for further details.

coefficient in Equation 5 is freely estimated for Y_2). This freely estimated carryover effect allows for the possibility that, for example, taking the medication on Day 1 has a cognitive enhancement effect that persists to some extent until Day 2. In all three groups, the threat factor (F_T) affects performance on the second measurement but not the first and is therefore reflective of a sequence/time effect. As in the Between \times Within design, this threat factor absorbs reactive effects, history effects, maturation effects, and regression to the mean. With this design, all terms in the comparison condition performance (F_c), causal effect (F_Δ), and extraneous variable (F_T) factor variance–covariance matrix ($\sigma_{F_c}^2$, $\sigma_{F_\Delta}^2$, $\sigma_{F_T}^2$, $\sigma_{c,\Delta}$, $\sigma_{c,T}$, and $\sigma_{\Delta,T}$) are identified. To make this concrete, a researcher would be able to estimate the correlation between sugar-pill performance and the cognitive enhancement effect, the correlation between sugar pill performance and the net effect of extraneous variables, the correlation between the net effect of extraneous variables and the cognitive enhancement effect, and the means and variances of sugar pill performance, the cognitive enhancement effect, and the net effect of extraneous variables. This is the first design discussed in this article to enable identification of all of these parameters.

Three-Group Nonrepeated Measures Design

This design is the same basic design as the three-group repeated measure design; however, it ensures that the same method of measurement is never repeated. As described earlier, all three groups are measured twice, but here different measures/test forms are used for each of the two measurements. That this design includes one group in which the comparison condition measurement is made first with Test Form A and one group in which the comparison condition measurement is made first with Test Form B effectively results in the equating of the different test forms by way

of the randomization process, and the experimental outcomes can therefore be considered calibrated to a common metric.

As in the three-group repeated measures design, this design allows for estimation of sequence- and time-related influences; however, because retesting occurs with novel methods/materials of measurement, such influences are potentially reduced. A three-group path-diagram representation of the application of Equation 5 to the data produced by the three-group nonrepeated measure design is depicted in Figure 5. The subscripts on Y correspond to the first and second measurement, with factor loadings and test difficulties varying according to the test form used. Standard factor identification constraints are applied, in this case by constraining the factor loading of Test Form A to 1 and constraining the intercept of Test Form A to 0. As in previous designs, extraneous variable factor, F_T , is reflective of reactive effects, history effects, maturation effects, and regression to the mean, and all terms in the comparison condition performance (F_c), causal effect (F_Δ), and extraneous variable (F_T) factor variance–covariance matrix ($\sigma_{F_c}^2$, $\sigma_{F_\Delta}^2$, $\sigma_{F_T}^2$, $\sigma_{c,\Delta}$, $\sigma_{c,T}$, and $\sigma_{\Delta,T}$) are identified.

It is of note that this design does not include all possible combinations of test form, measurement sequence, and condition (manipulation condition vs. comparison condition). Instead, the minimum number of combinations are included that allow for complete identification of the F_c , F_Δ , and F_T variance–covariance matrix. To illustrate, Test Forms A and B each appear in first (Y_1) and second (Y_2) positions in the sequence, as do manipulation-present and manipulation absent conditions; however, the manipulation-present condition is always paired with Test Form A. Of course, a fully counterbalanced, albeit much more complex, design that included all possible combinations of test form, measurement sequence, and condition would allow for assumptions regarding measurement invariance to be tested or, put another way, for would allow for testing of whether the causal effect depends on the

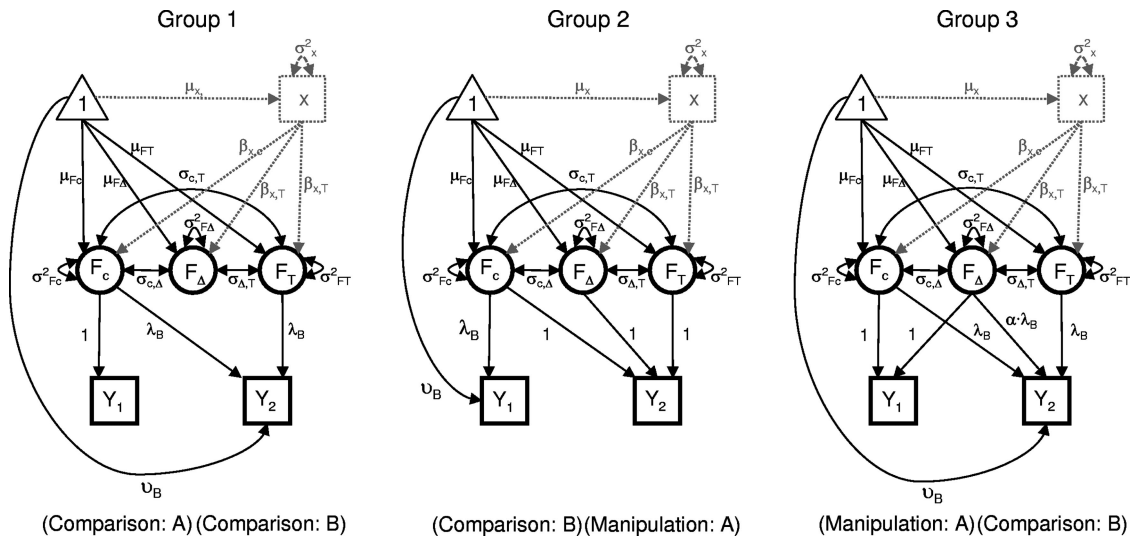


Figure 5. Structural equation model for the three-group nonrepeated measures design. The subscripts on Y correspond to the first measurement (1) and the second measurement (2). See Table 4 for a schematization of how data are collected for this design, Table 2 for a glossary of symbols used, and the in-text description for further details.

type of material used. This issue is discussed in further detail under the Assumptions and Limitations section of the Discussion (see the Measurement Invariance and Statistical Additivity section).

Simulation Studies

Method

Here, simulation is used to demonstrate how each of the described designs performs under a series of conditions in which potential threats to internal validity are progressively added. The strengths and weaknesses of each of the structural equation model–design pairings with respect to internal validity have already been discussed. This section serves to illustrate these strengths and weaknesses with actual numbers.

The simulations were specified to resemble the hypothetical cognitive enhancement experiment that has been used as an example throughout this article. In the comparison condition, participants take a sugar pill and are then administered a reasoning test. In the hypothetical manipulation condition, participants take a stimulant medication and are then administered a reasoning test, for which up to three alternate forms are available (i.e., each form is composed of the same types of questions representative of the same underlying ability but the different test forms do not contain any of the exact same questions). Scores on the reasoning tests were placed on continuous 0 through 15 point scales. In all generating models, true comparison (sugar pill) condition performance, F_c , was specified to have a mean (μ_{F_c}) of 7 and a variance ($\sigma_{F_c}^2$) of 1. Moreover the causal effect (the *cognitive enhancement effect*), F_Δ , was given a mean (μ_{F_Δ}) of 2 and a variance ($\sigma_{F_\Delta}^2$) of 1. In other words, the medication enhanced reasoning performance by 2 points on average, but this enhancement varied from person to person, such that, for example, some people's scores were enhanced by 1 point and others' were enhanced by 3 points. A small magnitude positive covariance ($\sigma_{c,\Delta}$) of .20 ($r = .20$) was set between comparison condition performance and the causal effect. An exogenous covariate, x (e.g., age), was also included. It was specified to have a variance (σ_x^2) of 1, and covariances with both comparison condition performance ($\sigma_{x,c}$) and the causal effect ($\sigma_{x,\Delta}$) of .40 ($r = .40$).

A best-case-scenario no-threat baseline simulation was first conducted and threats to validity were progressively added in four discrete steps. In Step 1, nontrivial error of measurement ($\sigma_u^2 = .20$) was specified. In Step 2, the designs that implement multiple test forms were specified to use test forms that were nonparallel ($\lambda_A = 1.00$, $\lambda_B = 1.10$, $\lambda_D = .80$, $v_A = 0.00$, $v_B = -1.00$, $v_D = 2.00$). In Step 3, a sequence effect was introduced ($\mu_T = 1$, $\sigma_T^2 = .50$). In Step 4, the sequence effect was specified to have nonzero covariances with comparison condition performance (F_c), the causal effect (F_m), and the covariate (x), such that $\sigma_{c,T} = .30$, $\sigma_{m,T} = .30$, and $\sigma_{x,T} = .30$. For all simulations, data were generated for a total of 200 hypothetical participants evenly distributed across groups. For each design at each step of the simulation, 100 data sets were generated and analyzed (i.e., each parameter estimate reported later is the average estimate from 100 replications).

In addition to those designs discussed earlier in this article, two other designs were fit to the simulated data. The first (a *counter-*

balanced order approach) was a within-subjects design in which the order of manipulation-present and manipulation-absent measurements is randomly counterbalanced between participants, data are collapsed across groups, and manipulation-present minus manipulation-absent difference scores are calculated for each individual and analyzed according to a conventional within-subject procedure in which dummy-coded variables representative of order (0 = *first*, 1 = *second*) are controlled for. The second (a *counterbalanced forms approach*) was a similar design in which different testing materials are used for each measurement, the pairing of testing materials with manipulation presence versus absence is counterbalanced between participants, data are collapsed across groups, and manipulation-present minus manipulation-absent difference scores are calculated for each individual and analyzed according to a conventional within-subject procedure in which dummy-coded variables representative of testing material (0 = *Test Form A*, 1 = *Test Form B*) are controlled for. Both designs are schematized in the bottom portion of Table 1. These two designs were fit because they might intuitively appear to control for threats associated with sequence effects (reactivity, maturation, and history threats) or noncomparable test forms (instrumentation threats), respectively. Note that although it would be possible to estimate a (somewhat constrained) model that includes a random threat factor (i.e., F_T) from data generated by the counterbalanced position design, this was not done here, because analyses of the counterbalanced position design are meant to serve as an illustration of the results of best current practice.

Results

Results of the simulations are presented in Table 4, which is subdivided into sections corresponding to the sequential steps described earlier. At the top row of each section, the true parameter values from the generating model are provided. In the ensuing rows, the average parameter estimates from 100 replications (with 200 participants per replication) for each design are provided. Average estimates that depart from the true values by more than .05 units are in bold. Because the true variances of comparison condition performance (F_c) and the manipulation effect (F_Δ) were set at 1 ($\sigma_{F_c}^2 = 1$, $\sigma_{F_\Delta}^2 = 1$), an estimate–true value discrepancy of .05 corresponds to a Cohen's d of .05, with respect to means (μ s), and a correlation unit of .05, with respect to covariances. Here, this .05 level is considered nontrivial bias that suggests that a design may be inappropriate for dealing with the validity threat.⁷

Baseline simulation. It can be seen that all approaches performed perfectly with respect to mean estimates, and all but one approach performed perfectly with respect to variance/covariance estimates, in this best case scenario simulation. That is, all approaches produced estimates of means of comparison condition performance (F_c) and the causal effect (F_Δ), and all but one approach produced estimates of the variances of and covariances

⁷ Parameter bias is sometimes indexed as a percentage deviation from the true parameter value, with bias greater than 5% being the conventional cutoff. Using percentages, however, is inappropriate when true parameter values are very small or 0. Nevertheless, the current .05 unit cutoff is compatible with the 5% convention in that F_Δ and F_c each have variances of 1, such that a .05 unit deviation is equivalent to a 5% deviation.

among F_c , F_Δ , and the covariate (x) that were nearly identical to those specified in the generating model. The only problematic design in this baseline simulation was the simple between-subjects design. Because this design does not have a provision for measuring the same participants in both manipulation and comparison conditions, the covariance between comparison condition performance (F_c) and the causal effect (F_Δ) cannot be estimated, which is equivalent to the $\sigma_{c,\Delta}$ parameter being incorrectly constrained to zero. This incorrect constraint produces a biased estimate of the variance of the causal effect ($\sigma_{F_\Delta}^2$). The discrepancy between the true and estimated values for $\sigma_{F_\Delta}^2$ is approximately .40 units, which is, not coincidentally, twice the value of the unmodeled $\sigma_{c,\Delta}$ covariance (see Appendix A for derivation). It is of note that, had the true value of $\sigma_{c,\Delta}$ been zero, the simple between-subjects design would have been well suited to (i.e., unbiased with respect to) these data. Even in the current situation, it accurately recovers the covariate–causal effect covariance, $\sigma_{x,\Delta}$.

Step 1: Imperfect measurement. The presence of measurement error produced a number of notable results. First, because no design, except for the common test-equating for experiments design, includes a measurement model that separates true (or common) variance from error (or unique) variance, it is not surprising that many of the estimates of the variance in comparison condition performance ($\sigma_{F_c}^2$) are inflated by the amount of unmodeled measurement error. This is typical in individual differences research and is generally considered tolerable when test reliabilities are moderate to high.

Second, in the basic within-subjects design and the two counterbalanced designs, the addition of measurement error resulted in an overestimate of the variance of the causal effect ($\sigma_{F_\Delta}^2$) and an underestimate of the covariance of comparison condition performance and the causal effect ($\sigma_{c,\Delta}$). It is illustrative to examine more closely the biases that arose in the simple within-subjects design. For this design, the estimate of the variance of the causal effect ($\sigma_{F_\Delta}^2$) is upwardly biased by .40 units, which is twice the amount of error associated with a single measurement. This is consistent with the well-known fact that, in calculating difference scores, the errors from both measurements become compounded (see, e.g., Cronbach & Furby, 1970). It can be seen that the estimate of the covariance between comparison condition performance and the causal effect ($\sigma_{c,\Delta}$) is biased downward by the value of the measurement error, which is consistent with a well-established literature on regression to the mean artifacts (Campbell & Kenny, 1999). These same results occur for the two counterbalanced approaches, which are, in this step, equivalent to the simple within-subjects design.

In contrast, measurement error did not bias estimates of the variance of the causal effect ($\sigma_{F_\Delta}^2$) or the comparison condition–causal effect covariance ($\sigma_{c,\Delta}$) in the test-equating for experiments design, the Between \times Within design, the three-group repeated measure design, or the three-group nonrepeated measures design. Why are the estimates from these designs not biased in ways similar to the ones discussed earlier? For the common test-equating for experiments design, the answer is straightforward. Measurement error does not affect estimates at the structural level because measurement error is removed at the measurement level. For the Between \times Within design, the three-group repeated measure design, and the three-group nonrepeated measures design, the answer is somewhat more novel. Because these designs each

include a control group that is measured multiple times in the absence of the experimental manipulation, these designs are able to quarantine measurement-error associated biases from the causal effect factor, F_Δ , and into the exogenous influences factor, F_T . That is, $\sigma_{c,\Delta}$ and $\sigma_{F_\Delta}^2$ parameters are estimated without bias, whereas the parameter representing the comparison condition–exogenous influences factor covariance ($\sigma_{c,T}$) is attenuated (by approximately .20 units, i.e., the magnitude of the measurement error), and the variance of the exogenous influences factor ($\sigma_{F_T}^2$) is inflated (by approximately .40 units, i.e., twice the measurement error). Although these latter parameters, which involve the exogenous influence factor, F_T , indeed depart from the values specified under the generating model, this is entirely acceptable, because F_T represents unwanted effects that, if not modeled, could bias estimates of the causal effect. That is, F_T does not represent phenomena of experimental interest but is rather included simply to decontaminate the causal effect factor, which does represent the phenomenon of interest.

Step 2: Nonparallel indicators. This step, in which alternate test forms were specified to be nonparallel, resulted in biased estimates of the variance of the causal effect (F_Δ) and the covariance between comparison condition performance and the causal effect ($\sigma_{c,\Delta}$) in the counterbalanced forms design but did not result in such biased estimates in the common test-equating for experiments design or the three-group nonrepeated measures design. The reasons for these differences are straightforward. The counterbalanced forms approach does not explicitly calibrate the different test forms to a common metric, whereas the common test-equating for experiments design and the three-group nonrepeated measures design do.

The results with respect to the counterbalanced forms design are somewhat concerning, given that researchers who use a counterbalanced forms approach likely do so because it might intuitively appear to correct for lack of measurement equivalence of alternate forms. The results with respect to the common test-equating for experiments design and the three-group nonrepeated measures design are, alternatively, encouraging. When the goal is to use a design that avoids the reactivity associated with repeated administrations of the same test, these latter two approaches each appear to be sensible choices.

Step 3: Sequence/time-related effects orthogonal to covariate and other components. This step, in which sequence/time-related effects were specified to occur, highlights the deficiencies of a number of designs. It is illustrative to first examine the biases that arose in the simple within-subjects design. This design produces a simple difference score representative of all experimental change, including both the causal effect of interest and the unwanted effects of history, maturation, reactivity, and regression to the mean. For example, the magnitude of the mean causal effect (μ_{F_Δ}) was overestimated at approximately 3.00, which is the sum of the mean of the causal effect (2.00) and the sequence/time-associated gain (1.00). Moreover, the variance in the causal effect ($\sigma_{F_\Delta}^2$) was estimated at 1.90, which reflects the sum of the actual between-person variation in the causal effect (1.00), the error terms from both measurements ($2 \times .20 = .40$), and the between-person variation in the unmodeled sequence/time-associated gain (.50).

The counterbalanced position approach avoided bias in the estimate of the mean causal effect (μ_{F_Δ}) but did not prevent the

bias in the estimate of the variance and covariance terms. It can be seen in Table 4 that estimated variances of comparison condition performance ($\sigma_{F_c}^2$) and manipulation-associated change ($\sigma_{F_{\Delta}}^2$) are highly inflated (1.45 compared with a true value of 1.00, and 1.87 compared with a true value of 1.00, respectively). Similarly, the covariance between comparison condition performance and manipulation-associated change ($\sigma_{c,\Delta}$) is dramatically underestimated ($-.24$ compared with a true value of $.20$). This is quite concerning given that researchers who use a counterbalanced position approach likely do so because it might intuitively appear that sequence/time effects should cancel out.

The common test-equating for experiments approach was also heavily biased by the addition of sequence/time-related effects that constituted this step. It can be seen in Table 5 that, for this approach, estimates of average comparison condition performance (μ_{F_c}), variance in comparison condition performance ($\sigma_{F_c}^2$), variance in the causal effect ($\sigma_{F_{\Delta}}^2$), and the covariance between comparison condition performance and the causal effect ($\sigma_{c,\Delta}$) were incrementally biased by these added specifications ($\mu_{F_c} = 7.99$ compared with a true value of 7.00; $\sigma_{F_c}^2 = 1.43$ compared with a true value of 1.00; $\sigma_{F_{\Delta}}^2 = .84$ compared with a true value of 1.00; and $\sigma_{c,\Delta} = .30$ compared with a true value of $.20$). Although it is likely that, in many cases, the common test-equating for experiments approach can be used to avoid reactive effects, if any sequence/time-related effects persist, the common test-equating for experiments approach is apparently ill-equipped to deal with them.

The approaches that were resilient to the specification of sequence/time-related effects that occurred in this step were those that explicitly modeled an F_T threat factor. These were the Between \times Within design, the three-group repeated measure design, and the three-group nonrepeated measures design. It can be seen in Table 4 that each of these three approaches produced accurate estimates of the mean of, variance of, and covariances involving the manipulation factor, F_{Δ} , and that the mean and variance of the sequence/time-related effects were appropriately absorbed by the extraneous variable factor, F_T .

Step 4: Sequence/time-related effects correlated with covariate and other components. This final step, in which the sequence/time-related effects were specified to be correlated with the covariate (x) and the comparison condition (F_c) and causal effect (F_{Δ}) factors, can be considered a worst case scenario and, accordingly, resulted in the most pervasive pattern of parameter biases. The simple within-subjects design is once again illustrative. It can be seen in Table 4 that, for this design, the unmodeled covariance between the covariate and the sequence effect ($\sigma_{x,T}$) was inappropriately absorbed into the covariance between the covariate and the causal effect ($\sigma_{x,\Delta}$), thereby inflating it. Similarly, the unmodeled covariance between comparison condition performance and the sequence effect ($\sigma_{c,T}$) was inappropriately absorbed into the estimate of the covariance between comparison condition performance and the causal effect, $\sigma_{c,\Delta}$.

Because it performed very well in all previous steps, the most notable bias arising in this step is observed in the Between \times Within approach. The inability of this approach to estimate the covariance between the causal effect and the sequence effect ($\sigma_{\Delta,T}$) resulted in bias in the estimated variance of the causal effect ($\sigma_{F_{\Delta}}^2$) by approximately twice the unmodeled $\sigma_{\Delta,T}$ term (see Appendix B for a derivation).

In this final step, the two advanced experimental designs introduced in this article—the three-group repeated measure design and the three-group nonrepeated measures design—remained resilient to the validity threats. All mean, variance, and covariance patterns involving the causal effect (F_{Δ}) remained unbiased. All threats, including sequence/time-related effects and regression to the mean, were absorbed by the threat factor (F_T), and for the three-group nonrepeated measures approach, the use of nonparallel alternate forms did not introduce estimation bias. These results illustrate the added value of the novel three-group repeated measure and the three-group nonrepeated measures approaches.

Discussion

Summary. In this article, a framework for collecting and analyzing data in randomized single-manipulation experiments was introduced. Researchers can vary the key manipulation, the instruments of measurement, and the sequences of the measurements and manipulations across participants, thus allowing both means and individual differences in the effects of each of these components to be statistically separated. A number of classical designs, a test-equating for experiments approach, and two advanced experimental designs were explicated and evaluated for their robustness to internal validity threats. Simulation studies illustrated that, although classical designs produce accurate estimates of mean effects, more sophisticated designs are often necessary for accurate inferences with respect to individual differences. Compared with the classical designs, the three-group repeated measure design and the three-group nonrepeated measures design both have particular advantages in their robustness to estimation bias when reactive, history, or maturation effects operate, particularly if individual differences in these effects covary with individual differences in the causal effect. Researchers, however, should not feel limited to the designs discussed in this article. The designs discussed should merely be taken as examples of how multiple-group structural equation models can be used to aid in the conceptualization of issues of individual differences in causal effects when designing experiments and analyzing data. Using the framework introduced in this article, researchers can customize their designs to fit their specific empirical needs.

Application and implementation of methods. The structural equation models described in this article can be implemented with any standard structural equation modeling software that allows for multiple group models. Example Mplus (L. K. Muthén & Muthén, 1998–2007) scripts for the Monte Carlo simulations reported here are available in the online supplemental materials for this article. These scripts may be advantageous for substantive researchers who are interested in producing power estimates when planning experiments, examining the feasibility of adaptations of the models discussed here, or analyzing real data that they have collected. It should be emphasized, however, that the Mplus software program is not necessary for carrying out the methods described here. Any contemporary structural equation program can be used to implement these methods.

Assumptions and limitations.

Convergent validity. Two of the designs introduced in this article are based on the assumption that experimental outcomes occur on unobserved factors that can be measured and operationalized in many alternative ways. Compared with those who are

Table 5
Results of Simulation Studies

Design	Means			Variances			Covariances					
	μ_{Fc}	$\mu_{F\Delta}$	μ_{FT}	σ_{Fc}^2	$\sigma_{F\Delta}^2$	σ_{FT}^2	$\sigma_{c,\Delta}$	$\sigma_{c,T}$	$\sigma_{\Delta,T}$	$\sigma_{x,c}$	$\sigma_{x,\Delta}$	$\sigma_{x,T}$
Baseline: Measures approach perfect reliability (reliability = .99)												
True	7.00	2.00	0.00	1.00	1.00	0.00	0.20	0.00	0.00	0.40	0.40	0.00
Simple within subjects	7.00	1.99	—	1.01	0.98	—	0.22	—	—	0.40	0.41	—
Simple between subjects	6.98	2.00	—	1.00	1.42	—	—	—	—	0.41	0.41	—
Between × Within	7.00	2.02	-0.00	1.02	1.01	0.02	0.23	-0.01	—	0.41	0.42	-0.00
Counterbalanced position	7.01	1.98	—	1.02	1.02	—	0.20	—	—	0.40	0.41	—
Counterbalanced forms	7.01	1.98	—	1.02	1.02	—	0.20	—	—	0.40	0.41	—
Common test equating for experiments	7.00	2.00	—	1.00	1.01	—	0.20	—	—	0.40	0.39	—
Three-group repeated measure	7.00	2.01	0.00	1.03	1.00	0.02	0.21	-0.01	-0.00	0.41	0.41	-0.00
Three-group nonrepeated measures	7.00	2.01	-0.00	1.03	0.99	0.02	0.21	-0.00	-0.00	0.41	0.41	0.00
Step 1: Imperfect measurement (reliability = .83)												
True	7.00	2.00	0.00	1.00	1.00	0.00	0.20	0.00	0.00	0.40	0.40	0.00
Simple within subjects	7.00	2.01	—	1.21	1.40	—	0.01	—	—	0.40	0.41	—
Simple between subjects	6.98	2.00	—	1.19	1.42	—	—	—	—	0.41	0.41	—
Between × Within	7.00	2.01	-0.01	1.18	1.00	0.39	0.20	-0.19	—	0.39	0.41	-0.00
Counterbalanced position	7.00	1.98	—	1.21	1.40	—	0.01	—	—	0.40	0.41	—
Counterbalanced forms	7.01	1.98	—	1.21	1.40	—	0.01	—	—	0.40	0.41	—
Common test equating for experiments	7.01	2.02	—	1.00	1.02	—	0.21	—	—	0.40	0.39	—
Three-group repeated measure	7.00	2.01	0.00	1.22	0.99	0.40	0.21	-0.20	0.00	0.40	0.41	0.00
Three-group nonrepeated measures	7.00	2.01	-0.00	1.22	1.01	0.40	0.20	-0.19	-0.00	0.40	0.41	0.00
Step 2: Nonparallel indicators												
True	7.00	2.00	0.00	1.00	1.00	0.00	0.20	0.00	0.00	0.40	0.40	0.00
Simple within subjects ^a												
Simple between subjects ^a												
Between × Within ^a												
Counterbalanced position ^a												
Counterbalanced forms	7.01	1.98	—	1.31	1.51	—	0.03	—	—	0.42	0.43	—
Common test equating for experiments	6.99	2.00	—	1.00	1.02	—	0.21	—	—	0.40	0.39	—
Three-group repeated measure ^a												
Three-group nonrepeated measures	7.01	2.01	0.00	1.18	1.01	0.36	0.18	-0.17	-0.01	0.40	0.39	0.01
Step 3: Sequence/reactive effects (orthogonal to x , F_c , and F_Δ)												
True	7.00	2.00	1.00	1.00	1.00	0.50	0.20	0.00	0.00	0.40	0.40	0.00
Simple within subjects	7.00	3.01	—	1.21	1.90	—	0.01	—	—	0.40	0.41	—
Simple between subjects ^a												
Between × Within	7.01	1.99	0.99	1.19	0.96	0.93	0.21	-0.22	—	0.40	0.40	0.00
Counterbalanced position	6.99	2.04	—	1.45	1.87	—	-0.24	—	—	0.41	0.41	—
Counterbalanced forms	7.01	2.98	—	1.31	2.07	—	0.03	—	—	0.43	0.43	—
Common test equating for experiments	7.99	2.01	—	1.43	0.84	—	0.30	—	—	0.43	0.36	—
Three-group repeated measure	6.99	2.00	0.99	1.17	1.00	0.90	0.19	-0.19	0.01	0.40	0.40	0.01
Three-group nonrepeated measures	7.00	2.00	1.01	1.18	1.02	0.87	0.18	-0.17	-0.02	0.40	0.39	0.00

(table continues)

Table 5 (continued)

Design	Means			Variances			Covariances					
	μ_{F_c}	μ_{F_Δ}	μ_{F_T}	$\sigma_{F_c}^2$	$\sigma_{F_\Delta}^2$	$\sigma_{F_T}^2$	$\sigma_{c,\Delta}$	$\sigma_{c,T}$	$\sigma_{\Delta,T}$	$\sigma_{x,c}$	$\sigma_{x,\Delta}$	$\sigma_{x,T}$
Step 4: Sequence/reactive effects (correlated with x , F_c , and F_Δ)												
True	7.00	2.00	1.00	1.00	1.00	0.50	0.20	0.30	0.30	0.40	0.40	0.30
Simple within subjects	7.00	3.01	—	1.21	2.51	—	0.31	—	—	0.41	0.72	—
Simple between subjects ^a												
Between \times Within	7.00	2.03	1.00	1.21	1.63	0.89	0.26	0.09	—	0.41	0.42	0.30
Counterbalanced position	7.00	2.04	—	1.76	1.87	—	-0.09	—	—	0.56	0.41	—
Counterbalanced forms	7.01	2.98	—	1.31	2.74	—	0.36	—	—	0.43	0.75	—
Common test equating for experiments	8.00	2.00	—	2.26	1.31	—	0.37	—	—	0.70	0.39	—
Three-group repeated measure	6.99	2.00	0.98	1.17	1.00	0.90	0.19	0.10	0.30	0.40	0.40	0.31
Three-group nonrepeated measures	7.00	2.01	1.01	1.18	1.01	0.87	0.19	0.12	0.28	0.40	0.39	0.30

Note. Rows labeled *True* contain the parameter values that were specified in the generating model. Values in bold depart from the true value by greater than .05 units. F_c is a factor representing comparison condition performance. F_Δ is a factor representing the manipulation-associated causal effect. F_T is a factor representing threats to experimental validity. A dash indicates that the parameter was not estimated, which is equivalent to it being fixed to zero.
^a The threat was not relevant to the design in question.

interested in general constructs, researchers who are interested in specific outcomes or behaviors may therefore find some designs less suited to their goals.

Measurement invariance and statistical additivity. The assumption that the outcomes in an experiment occur on general factors, rather than specific tests, requires fulfillment of measurement invariance across positions in the sequence (i.e., whether the measure was administered first or second) and measurements in the presence versus absence of the manipulation. If measurement invariance holds, it can be concluded that the changes occur on the factors rather than the specific measures. A full treatment of measurement invariance is beyond the scope of the current article, but a number of detailed articles on the subject exist (e.g., Horn & McArdle, 1992; Meredith, 1993).

The methods advocated here also rely on the closely related assumption of statistical additivity of the variance components and of the mean effects. That is, these methods are based on the premise that experimental situations in which none or a subset of the change influences (e.g., sequence and manipulations effects) are operating can be used to make inferences about experimental situations in which all influences are operating, such that the isolated components add together to form the total change. In some cases, examinations of measurement invariance can be used to test these assumptions, but in other cases, more elaborate procedural and statistical methods may be required. One design that was not discussed here but can be used to investigate whether the effects of an experimental manipulation differ according to whether individuals were previously tested is the Solomon four-group design (Solomon, 1949). An analogous design could prove useful in determining whether the effects of an experimental manipulation differ according to whether individuals were previously tested on the same or a different version of a test or measure (these possibilities are closely related to what Poulton, 1975, termed *range effects*). For examination of issues of nonadditivity of unmeasured variance components (latent factors), new developments in nonlin-

ear and interactive factor analysis are likely to prove useful (e.g., Klein & Moosbrugger, 2000; Tucker-Drob, 2009).

Covariation and causation. The maxim that covariation does not (necessarily) imply causation holds true for the individual differences approaches described in this article. Just because an exogenous individual differences variable is related to the size of the causal effect of an experimental manipulation does not mean that this variable causes a resistance or vulnerability to the experimental manipulation. It is very possible that an unmeasured set of third variables plays the true causal role. However, the methods reviewed here can be used to isolate individual differences in the causal effect of the manipulation from those related to other characteristics of the experimental situation. Such assignment of variation to the appropriate sources is an important preliminary step in the falsification of causal hypotheses. Of course, in instances in which the exogenous correlate of the causal effect is itself a manipulated variable, this caveat does not apply.

Focus on the causal effect. The methods described in this article were developed with the goal of distilling mean and individual differences associated with the causal effect from those associated with other aspects of the experimental situation, such as reactivity, maturation, and history. These methods may therefore be less useful when the primary research focus is on (what are, in the current context, considered) validity threats. For example, the cognitive psychologist may be interested in retest-related transfer effects, the developmental psychologist may be interested in age-related maturation, and the demographer may be interested in history-related cohort effects. In the current framework, these influences are all considered sequence-related threats, and provisions are not included for separating them from one another. Other methodological works specifically focus on these sorts of issues (e.g., Baltes & Nesselroade, 1970; McArdle & Woodcock, 1997), and the interested reader is encouraged to consult them.

Future Directions

The approaches reviewed and introduced in this article can be directly implemented for experimental research in many substantive areas. Nevertheless, there is much room for future work. One main issue is power. Power, of course, depends on a host of characteristics of the sample, data pattern, and analytical model, such that any power study will therefore be limited in its generalizability. In designing a specific experiment, the most appropriate type of power analysis would therefore be one tailored to that experiment, but a general treatment of power for individual-differences approaches to experiments would nevertheless be quite useful.

Although the framework put forth is quite general for many sorts of single-manipulation experiments, a number of extensions are warranted. Perhaps the most obvious extension involves the addition of provisions for multiple levels of the manipulation and three or more measurements per person. This would allow for conversion of the statistical framework from one of a difference score approach to one of a growth curve, or random effects, approach. One way that this could be achieved is by allowing the m and p coefficients in Equation 5 to act as growth-curve basis coefficients (see, e.g., McArdle & Nesselrode, 2003), taking on values as parametric, or freely estimated, functions of manipulation level and occasion of measurement, respectively. Extension of the framework to multiple manipulation experiments would also be particularly valuable. Such an extension would require the development of provisions for a host of added methodological issues, including interference and interactions among the different manipulations. Finally, as discussed earlier, future provisions for latent variable interactions in experiments would be particularly valuable.

Conclusion

This article focused on three core ideas. First, random assignment, the sine qua non of experimental science, permits researchers to examine not only the average effects of a manipulation or treatment but also individual differences in responsiveness to the manipulation/treatment and their correlates. When group membership is randomly assigned and an experimental manipulation is applied to participants in only one of the two groups, any differences between the groups can be attributed to the presence versus absence of the manipulation, including, of course, mean differences but also any differences in variances, covariances, and regression relationships. Second, individual differences are routinely neglected in experimental science, in part because researchers lack appropriate experimental designs and data analytical strategies. This article begins to fill this gap in the methodological literature by presenting novel individual-differences-oriented approaches to experimental design and data analysis that control for threats to internal validity by way of integration of classical within-subjects, between-subjects, and test-equating methods. Third, individual differences in the causal effects of experimental manipulations and the relations between individual causal effects and person characteristics are, rather than being noise or nuisance phenomena, critically important concepts for both basic theory and applied psychology.

References

- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington DC: American Council on Education.
- Baltes, P. B., & Nesselrode, J. R. (1970). Multivariate longitudinal and cross-sectional sequences for analyzing ontogenetic and generational change: A methodological note. *Developmental Psychology, 2*, 163–168. doi:10.1037/h0028743
- Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11*, 142–163. doi:10.1037/1082-989X.11.2.142
- Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin, 104*, 396–404. doi:10.1037/0033-2909.104.3.396
- Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York, NY: Guilford Press.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.
- Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426–443. doi:10.1037/h0026714
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Harcourt Brace Jovanovich College.
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671–684. doi:10.1037/h0043943
- Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116–127. doi:10.1037/h0076829
- Cronbach, L. J., & Furby, L. (1970). How we should measure “change”—or should we? *Psychological Bulletin, 74*, 68–80. doi:10.1037/h0029382
- DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3*, 412–423. doi:10.1037/1082-989X.3.4.412
- Fisher, R. A. (1925). *Statistical methods for research workers*. London, England: Oliver & Boyd.
- Glynn, A. N. (2010). *The product and difference fallacies for indirect effects*. Unpublished manuscript.
- Greely, H., Sahakian, B., Harris, J., Kessler, R. C., Gazzaniga, M., Campbell, P., & Farah, M. J. (2008). Towards responsible use of cognitive-enhancing drugs by the healthy. *Nature, 456*, 702–705. doi:10.1038/456702a
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960. doi:10.2307/2289064
- Hom, J. L., & McArdle, J. J. (1992). A practical guide to measurement invariance in research on aging. *Experimental Aging Research, 18*, 117–144.
- Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-participant designs. *Psychological Methods, 6*, 115–134. doi:10.1037/1082-989X.6.2.115
- Judd, C. M., McClelland, G. H., & Smith, E. R. (1996). Testing treatment by covariate interactions when treatment varies within subjects. *Psychological Methods, 1*, 366–378. doi:10.1037/1082-989X.1.4.366
- Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods, 8*, 115–128. doi:10.1037/1082-989X.8.2.115
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.
- Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65*, 457–474. doi:10.1007/BF02296338
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.
- Masters, G. N. (1985). Common-person equating with the Rasch model. *Applied Psychological Measurement, 9*, 73–82. doi:10.1177/014662168500900107

- McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological contributions* (pp. 223–267). Hillsdale, NJ: Erlbaum.
- McArdle, J. J., & Nesselroade, J. R. (2003). Growth curve analysis in contemporary psychological research. In J. Schinka & W. Velicer (Eds.), *Comprehensive handbook of psychology: Volume II. Research methods in psychology* (pp. 447–480). New York, NY: Pergamon Press.
- McArdle, J. J., & Woodcock, J. R. (1997). Expanding test–rest designs to include developmental time-lag components. *Psychological Methods*, 2, 403–435. doi:10.1037/1082-989X.2.4.403
- McCall, W. A. (1923). *How to experiment in education*. New York, NY: Macmillan.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543. doi:10.1007/BF02294825
- Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, 2, 371–402. doi:10.1037/1082-989X.2.4.371
- Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.
- Poulton, E. C. (1975). Range effects in experiments on people. *American Journal of Psychology*, 88, 3–32. doi:10.2307/1421662
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100, 322–331. doi:10.1198/016214504000001880
- Salthouse, T. A. (2007). Implications of within-person variability in cognitive and neuropsychological functioning for the interpretation of change. *Neuropsychology*, 21, 401–411. doi:10.1037/0894-4105.21.4.401
- Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology*, 22, 800–811. doi:10.1037/a0013091
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.
- Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, 46, 137–150. doi:10.1037/h0062958
- Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. *Psychometrika*, 43, 381–396. doi:10.1007/BF02293647
- Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1, 39–54.
- Steyer, R., Nachtigall, C., Wüthrich-Martone, O., & Kraus, K. (2002). Causal regression models III: Covariates, conditional and unconditional average causal effects. *Methods of Psychological Research Online*, 7, 41–68.
- Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the lifespan. *Developmental Psychology*, 45, 1097–1118. doi:10.1037/a0015864

Appendix A

Biased Variance of Manipulation Estimate in Simple Between-Subjects Design

In the simple between-subjects design, comparison condition performance, F_c , and the causal effect, F_Δ , are assumed to be uncorrelated for identification purposes. Here, it is shown how this assumption, if violated, can bias the estimated variance of the causal effect.

The value of the simple between-subjects design derives from the premise that any differences in means, variances, and covariances observed between control and experimental groups can be attributed to the effect of the manipulation. One can represent this more formally as follows:

$$Y^{\text{Group}1} = F_c + u, \text{ and} \quad (\text{A1})$$

$$Y^{\text{Group}2} = F_c + F_\Delta + u, \quad (\text{A2})$$

where Y is the measured outcome, and its superscript denotes the randomly assigned group. F_c is comparison condition performance and is allowed to have a mean, μ_{F_c} , and a variance, $\sigma_{F_c}^2$. The causal effect, F_Δ , is similarly allowed to have a mean, μ_{F_Δ} , and a variance, $\sigma_{F_\Delta}^2$. For identification purposes, measurement error, u , is not allowed in this model but in reality may have variance σ_u^2 . Also for identification purposes, no covariance ($\sigma_{c,\Delta}$) is allowed between F_c and F_Δ , although one may exist in reality. It follows that the variances of Y in Groups 1 and 2 are actually

$$\sigma_Y^{2\text{Group}1} = \sigma_{F_c}^2 + \sigma_u^2, \text{ and} \quad (\text{A3})$$

$$\sigma_Y^{2\text{Group}2} = \sigma_{F_c}^2 + \sigma_{F_\Delta}^2 + 2 \times \sigma_{c,\Delta} + \sigma_u^2, \quad (\text{A4})$$

but are modeled as

$$\sigma_Y^{2\text{Group}1} = \hat{\sigma}_{F_c}^2, \text{ and} \quad (\text{A5})$$

$$\sigma_Y^{2\text{Group}2} = \hat{\sigma}_{F_c}^2 + \hat{\sigma}_{F_\Delta}^2. \quad (\text{A6})$$

Subtracting A5 from A6 yields the predicted variance of the causal effect, $\hat{\sigma}_{F_\Delta}^2$,

$$\hat{\sigma}_{F_\Delta}^2 = \sigma_Y^{2\text{Group}2} - \sigma_Y^{2\text{Group}1}. \quad (\text{A7})$$

Substituting A3 and A4 into A7 yields

$$\hat{\sigma}_{F_\Delta}^2 = (\sigma_{F_c}^2 + \sigma_{F_\Delta}^2 + 2 \times \sigma_{c,\Delta} + \sigma_u^2) - (\sigma_{F_c}^2 + \sigma_u^2), \quad (\text{A8})$$

which reduces to

$$\hat{\sigma}_{F_\Delta}^2 = \sigma_{F_\Delta}^2 + 2 \times \sigma_{c,\Delta}. \quad (\text{A9})$$

Equation A9 shows that $\hat{\sigma}_{F_\Delta}^2$ will be biased by twice the covariance between comparison condition performance and the causal effect. If $\sigma_{c,\Delta}$ is positive, $\hat{\sigma}_{F_\Delta}^2$ will be inflated, whereas if $\sigma_{c,\Delta}$ is negative, $\hat{\sigma}_{F_\Delta}^2$ will be attenuated.

Appendix B

Biased Variance of Manipulation Estimate in the Between \times Within Design

In the Between \times Within design, the causal effect, F_{Δ} , and the threat-related change, F_T , are assumed to be uncorrelated for identification purposes. Here, it is shown how this assumption, if violated, can bias the estimated variance of the causal effect.

The Between \times Within design can be written as

$$Y_1 = F_c + u_1, \quad (\text{B1})$$

$$Y_2^{\text{Group1}} = F_c + F_T + u_2, \quad (\text{B2})$$

$$Y_2^{\text{Group2}} = F_c + F_T + F_{\Delta} + u_2, \quad (\text{B3})$$

where Y is the measured outcome, its superscript denotes the randomly assigned group, and its subscript denotes whether it is the first or second measurement. F_c is allowed to have a mean, μ_{F_c} , and a variance, $\sigma_{F_c}^2$. F_{Δ} is the causal effect and is similarly allowed to have a mean, $\mu_{F_{\Delta}}$, and a variance, $\sigma_{F_{\Delta}}^2$. F_T is the threat-related factor and is allowed to have a mean, μ_{F_T} , and a variance, $\sigma_{F_T}^2$. The covariances between F_c and F_{Δ} ($\sigma_{c,\Delta}$) and between F_c and F_T ($\sigma_{c,T}$) are allowed, but to achieve identification, the covariance between F_{Δ} and F_T ($\sigma_{\Delta,T}$) is not allowed, although it may exist in reality. Similarly, for identification purposes, measurement error, u , is not identified in this model but in reality may have a variance of σ_u^2 . It follows that the variances and covariances of Y_1 and Y_2 in Groups 1 and 2 are actually as follows:

$$\sigma_{Y_1}^2 = \sigma_{F_c}^2 + \sigma_u^2, \quad (\text{B4})$$

$$\sigma_{Y_2}^{2\text{Group1}} = \sigma_{F_c}^2 + \sigma_{F_T}^2 + 2 \times \sigma_{c,T} + \sigma_u^2, \quad (\text{B5})$$

$$\sigma_{Y_2}^{2\text{Group2}} = \sigma_{F_c}^2 + \sigma_{F_T}^2 + \sigma_{F_{\Delta}}^2 + 2 \times (\sigma_{c,\Delta} + \sigma_{c,T} + \sigma_{\Delta,T}) + \sigma_u^2, \quad (\text{B6})$$

$$\sigma_{Y_1, Y_2}^{\text{Group1}} = \sigma_{F_c}^2 + \sigma_{c,T}, \text{ and} \quad (\text{B7})$$

$$\sigma_{Y_1, Y_2}^{\text{Group2}} = \sigma_{F_c}^2 + \sigma_{c,T} + \sigma_{c,\Delta}, \quad (\text{B8})$$

but are modeled as

$$\sigma_{Y_1}^2 = \hat{\sigma}_{F_c}^2, \quad (\text{B9})$$

$$\sigma_{Y_2}^{2\text{Group1}} = \hat{\sigma}_{F_c}^2 + \hat{\sigma}_{F_T}^2 + 2 \times \hat{\sigma}_{c,T}, \quad (\text{B10})$$

$$\sigma_{Y_2}^{2\text{Group2}} = \hat{\sigma}_{F_c}^2 + \hat{\sigma}_{F_T}^2 + \hat{\sigma}_{F_{\Delta}}^2 + 2 \times (\hat{\sigma}_{c,\Delta} + \hat{\sigma}_{c,T}), \quad (\text{B11})$$

$$\sigma_{Y_1, Y_2}^{\text{Group1}} = \hat{\sigma}_{F_c}^2 + \hat{\sigma}_{c,T}, \text{ and} \quad (\text{B12})$$

$$\sigma_{Y_1, Y_2}^{\text{Group2}} = \hat{\sigma}_{F_c}^2 + \hat{\sigma}_{c,T} + \hat{\sigma}_{c,\Delta}. \quad (\text{B13})$$

Subtracting B12 from B13 yields

$$\hat{\sigma}_{c,\Delta} = \sigma_{Y_1, Y_2}^{\text{Group2}} - \sigma_{Y_1, Y_2}^{\text{Group1}}, \quad (\text{B14})$$

and substituting B7 and B8 into B14 yields

$$\hat{\sigma}_{c,\Delta} = \sigma_{c,\Delta}. \quad (\text{B15})$$

Equation B15 demonstrates that unmodeled measurement error does not result in a biased estimate of the covariance of comparison condition performance and the causal effect. Where does the regression to the mean that might have been expected go? The answer comes from substituting B9, B4, and B7 into B12 and solving for $\hat{\sigma}_{c,T}$, which yields

$$\hat{\sigma}_{c,T} = \sigma_{c,T} - \sigma_u^2. \quad (\text{B16})$$

Equation B16 demonstrates that the regression to the mean induced by unmodeled measurement error is actually absorbed into the relation between comparison condition performance and the threat-related change. That is, $\hat{\sigma}_{c,T}$, rather than $\hat{\sigma}_{c,\Delta}$, is attenuated by an amount equal to the measurement error in Y .

Subtracting B10 from B11 yields

$$\sigma_{Y_2}^{2\text{Group2}} - \sigma_{Y_2}^{2\text{Group1}} = \hat{\sigma}_{F_{\Delta}}^2 + 2 \times \hat{\sigma}_{c,\Delta}. \quad (\text{B17})$$

Substituting B5 and B6 into B17 reduces to

$$\sigma_{F_{\Delta}}^2 + 2 \times \sigma_{c,\Delta} + 2 \times \sigma_{\Delta,T} = \hat{\sigma}_{F_{\Delta}}^2 + 2 \times \hat{\sigma}_{c,\Delta}, \quad (\text{B18})$$

and substituting B15 into B18 reduces to

$$\hat{\sigma}_{F_{\Delta}}^2 = \sigma_{F_{\Delta}}^2 + 2 \times \sigma_{\Delta,T}. \quad (\text{B19})$$

Equation B19 demonstrates that $\hat{\sigma}_{F_{\Delta}}^2$ will be biased by twice the covariance between the causal effect and the threat-related factor. If $\sigma_{\Delta,T}$ is positive, $\hat{\sigma}_{F_{\Delta}}^2$ will be inflated, whereas if $\sigma_{\Delta,T}$ is negative, $\hat{\sigma}_{F_{\Delta}}^2$ will be attenuated.

(Appendices continue)

Appendix C

Parameter Specifications for Application of Equation 5 to Various Experimental Designs

Group	w	m	p	λ_w	v_w	σ_w^2	μ_{Fc}	$\mu_{F\Delta}$	μ_{FT}	σ_{Fc}^2	$\sigma_{F\Delta}^2$	σ_{FT}^2	$\sigma_{c,\Delta}$	$\sigma_{c,T}$	$\sigma_{\Delta,T}$
Simple within subjects															
1 (1st measurement)	A	0	—	1	0	0	μ_{Fc}	$\mu_{F\Delta}$	—	σ_{Fc}^2	$\sigma_{F\Delta}^2$	—	$\sigma_{c,\Delta}$	—	—
1 (2nd measurement)	A	1	—	1	0	0									
Simple between subjects															
1	A	0	—				μ_{Fc}	$\mu_{F\Delta}$	—	σ_{Fc}^2	$\sigma_{F\Delta}^2$	—	0	—	—
2	A	1	—	1	0	0									
Between \times Within															
1 (1st measurement)	A	0	0	1	0	0	μ_{Fc}	$\mu_{F\Delta}$	μ_{FT}	σ_{Fc}^2	$\sigma_{F\Delta}^2$	σ_{FT}^2	$\sigma_{c,\Delta}$	$\sigma_{c,T}$	0
1 (2nd measurement)	A	0	1	1	0	0									
2 (1st measurement)	A	0	0	1	0	0									
2 (2nd measurement)	A	1	1	1	0	0									
Common test equating for experiments															
1 (Test A)	A	0					μ_{Fc}	$\mu_{F\Delta}$	—	σ_{Fc}^2	$\sigma_{F\Delta}^2$	—	$\sigma_{c,\Delta}$	—	—
1 (Test B)	B	1	—	λ_B	v_B	σ_B^2									
1 (Test D)	D	0	—	λ_D	v_D	σ_D^2									
2 (Test A)	A	1	—	1	0	σ_A^2									
2 (Test B)	B	0	—	λ_B	v_B	σ_B^2									
Three-group repeated measure															
1 (1st measurement)	A	0	0	1	0	0	μ_{Fc}	$\mu_{F\Delta}$	μ_{FT}	σ_{Fc}^2	$\sigma_{F\Delta}^2$	σ_{FT}^2	$\sigma_{c,\Delta}$	$\sigma_{c,T}$	$\sigma_{\Delta,T}$
1 (2nd measurement)	A	0	1	1	0	0									
2 (1st measurement)	A	0	0	1	0	0									
2 (2nd measurement)	A	1	1	1	0	0									
3 (1st measurement)	A	1	0	1	0	0									
3 (2nd measurement)	A	α	1	1	0	0									
Three-group nonrepeated measures															
1 (1st measurement)	A	0	0	1	0	0	μ_{Fc}	$\mu_{F\Delta}$	μ_{FT}	σ_{Fc}^2	$\sigma_{F\Delta}^2$	σ_{FT}^2	$\sigma_{c,\Delta}$	$\sigma_{c,T}$	$\sigma_{\Delta,T}$
1 (2nd measurement)	B	0	1	λ_B	v_B	0									
2 (1st measurement)	B	0	0	λ_B	v_B	0									
2 (2nd measurement)	A	1	1	1	0	0									
3 (1st measurement)	A	1	0	1	0	0									
3 (2nd measurement)	B	$\lambda_B \times \alpha$	1	λ_B	v_B	0									

Note. Parameters labeled 0 and 1 are fixed to those values. Other parameters are freely estimated. A dash denotes that a parameter corresponds to a factor that is not included in the model. Means, variances, and covariances of F_c , F_Δ , and F_T factors are constrained to equality across groups within each design.

Received October 6, 2008
 Revision received October 20, 2010
 Accepted November 11, 2010 ■