Some Issues Related to the Use of Randomized Trials in the Field of Program Evaluation

© 2011

By

John Arnold Gargani
Abstract

Some Issues Related to the Use of Randomized Trials in the Field of Program Evaluation

By

John Arnold Gargani

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

This dissertation explores three topics related to the use of randomized trials in the field of program evaluation. The dissertation is organized as three papers, each of which is intended to stand alone. The first paper addresses the problem that evaluators face when they must analyze data from imperfectly implemented randomized trials. The paper focuses on the two-group pretest-posttest design, widely used by evaluators to estimate the causal effects of programs. When this design is used, evaluators must make a choice between analyzing their data with the covariance adjustment approach or gain score approach. A great deal has been written about how this choice should be made, but it has largely been intended for researchers engaging in quasi-experimental or observational studies. This paper proposes a simple, objective way for evaluators to make this choice—the mixed approach—when, as is often the case, they find that an experiment has been weakened by real-world circumstances. When this occurs, evaluators may not be aware of the nature or extent of the flaws in their realized research designs. The mixed approach allows evaluators to choose between the gain score and covariance adjustment approaches more advantageously when (1) ceiling and floor effects are minimal, (2) it is reasonable to estimate an average treatment effect, and (3) the principal uncertainty is whether randomization has ensured the long-term comparability of groups or whether it failed in ways that led to the nonequivalence of groups at the time of the pretest.

The second paper describes a reparameterized Rasch model (RRM). Unlike a traditionally parameterized Rasch model, the RRM produces estimates of item group difficulties and associated tests of significance. Because the RRM is a Rasch model, the item difficulty estimates produced by a traditionally parameterized Rasch model can be calculated using the RRM. The RRM is compared to the traditionally parameterized Rasch model and the multidimensional random coefficients multinomial logit model. An example is provided in which all three models are fit as hierarchical generalized linear models, the models are formally compared, and the RRM is used to answer two research questions. The first question relates to the validity of reverse-scored items. Two groups of items were considered—items with original wording (left) and parallel items with presumed opposite wording (right). The RRM was used to determine whether the left item group and reverse-scored right item group had similar average item difficulties. The
second question relates to whether differences between the item groups were associated with respondents’ years of professional experience. To answer this question, the RRM was expanded to include a person-level main effect and cross-level (experience by item group) interaction. For evaluators implementing studies with experimental designs, the RRM provides a method of validating a measure while simultaneously answering substantive research questions.

The third paper describes a concrete process that stakeholders can use to make predictions about the future performance of programs in local contexts. Within the field of evaluation, the discussion of validity as it relates to outcome evaluation has been focused on questions of internal validity (Did it work?) to the exclusion of external validity (Will it work?). However, recent debates about the credibility of evaluation evidence have called attention to how evaluations can inform predictions about future performance. Using this as a starting point, I expand upon the traditional framework regarding external validity so closely associated with Donald Campbell. I argue that while there are three broad strategies for strengthening predictions of future program performance—design strategies, reporting strategies, and use strategies—only use strategies have the potential to substantially improve evaluation practice in the foreseeable future. Accordingly, the process I describe is one possible use strategy that is collaborative, systematic, feasible, and transparent.
# Table of Contents

**Paper 1**  
A Proposed Method for Choosing Between Gain Score and Covariance Adjustment Approaches When Estimating Average Treatment Effects With Degraded Experiments

**Paper 2**  
A Reparameterized Rasch Model

**Paper 3**  
External Validity, Evaluation Evidence, and Action in Local Contexts
Acknowledgements

A dissertation represents a substantial investment of time and effort. I would like to thank my wife, Mary Kathryn, for encouraging me over the years. Her loving support reminded me every day how lucky I am. I would also like to thank my children, Isabella, Enzo, and Elio, who were too young to fully understand what their father was doing but old enough to recognize that it was important to me. Their encouragement was unexpected, unbounded, and irrepressible. I can only hope to live up to the fine examples they have set for me.

Stewart Donaldson has been terrifically supportive and I cannot thank him enough. His friendship, advice, and comments have been invaluable and, more than once, ensured that I continued to forge ahead. I look forward to our growing collaboration on the many interests that we share. Michael Strong has impressed me tremendously for many reasons, not the least of which is that he has seen it all yet remains enthusiastic. His fascinating data and imaginative research have enlightened me, and his moral support helped me keep calm and carry on in the best British tradition.

I would like to thank my committee, Mark Wilson, Sophia Rabe-Hesketh, and Philip Tetlock, for their support. The only thing more difficult than writing a dissertation is reading one written by someone else. They did so repeatedly and I thank them for their efforts. I could not have done it without them.
Introduction

The two-group pretest-posttest design may be the most frequently used design in social science research (Cook & Campbell, 1979), and program evaluators routinely implement it in experimental and non-experimental settings to estimate the causal effects of programs. When they do, they face a choice between analyzing their data with either a gain score approach or a covariance adjustment approach. The former incorporates gain scores (the posttest score less the pretest score) as the outcome measure of a statistical model. Often, this is a linear regression with an independent variable that indicates the group to which participants belonged. The latter incorporates posttest scores as the outcome variable and, in the context of linear regression, includes pretest scores as a covariate in addition to the group indicator. The choice of approach has the potential to affect precision (Lord, 1956; Oakes & Feldman, 2001) and bias (Lord, 1967; Wainer, 1991), which has led a number of researchers to suggest rules for making an appropriate choice. However, in this work the research contexts—the statistical models, research purposes, and study designs for which the rules are applicable—have varied greatly and at times have been left vague (Maris, 1998). Moreover, this advice has not always been correct (Edwards, 2001; Rogosa & Willett, 1983; Zumbo, 1999). Evaluators and other social science researchers would therefore be well served by a simple, sound basis for choosing between approaches.

The purpose of this paper is to describe a simple rule that can be applied to linear regression models used to estimate a program’s average treatment effect with data from a randomized trial. When a randomized trial is well implemented, the choice is straightforward. However, evaluations are conducted in real-world settings that almost always degrade experimental designs to some extent. The full extent of that degradation and the degree to which it compromises results are usually unknown to the evaluator. The procedure described in this paper takes this into account and is intended for evaluators who must choose between approaches when they are uncertain of the quality of their realized experiment.

I begin by specifying the research context with which I am concerned, describe the ways in which past researchers have used precision and bias as criteria for choosing between approaches in various contexts, and then discuss why these past suggestions do not provide an adequate solution for the context under consideration. Then I present what I call the mixed approach, which uses an objective rule to select the approach that minimizes bias, maximizes statistical power, and maintains the nominal Type I error rate. I describe how I simulated data to benchmark the performance of the mixed approach, present the results of those simulations, and evaluate results. For the data structures considered, either (1) the covariance adjustment or gain score approach performed better, or (2) neither performed well. Throughout, the mixed approach tended to select the superior approach when there was one, and it did no worse when the there was not. However, results suggest that a small penalty in terms of bias, power, or Type I error rate may sometimes be the price of using the rule-based mixed approach.
Background

The choice between the gain score approach and the covariance adjustment approach has attracted considerable attention over the years. To clarify what will follow, I describe the research context that I address. This includes the purpose of the research (estimating the average treatment effect), the statistical models (ordinary least squares regression), and the research design (a degraded experiment). I also summarize several procedures that have been suggested for making this choice, each of which has its own strengths and weaknesses.

The Research Context

Purpose: Estimation of the Average Treatment Effect

The average treatment effect (ATE, Rosenbaum & Rubin, 1983) is the expected impact of a program on participants, and can be defined in terms of the Rubin causal model (RCM, Holland, 1986; Rubin, 1974, 2004). From this perspective, causal effects are comparisons of potential outcomes. The RCM takes a set of distinct conditions to which an individual could be exposed, and then for each condition maps exposure to the outcome that would be realized if exposed. For the two-group pretest-posttest design, a participant can be exposed to a treatment condition (1) or not (0). The potential outcomes for individual \( i \) that are associated with exposure to these two conditions are denoted \( Y_i(1) \) and \( Y_i(0) \), respectively. For example, \( Y_i(1) \) might be the one-year gain in reading ability that a particular student would realize if she were to attend a new reading program, and \( Y_i(0) \) is the quarter-year gain that she would realize if she were not to attend. The causal effect for an individual is defined as \( Y_i(1) - Y_i(0) \), so for this example the causal effect for the student is a net three-quarters-year gain.

The RCM depends on the stable-unit-treatment-value assumption (SUTVA), which states that that, regardless of how individuals are assigned to conditions, each individual has only one potential outcome per condition. This can be violated if, for example, the quality of program implementation depends on the number of participants who were recruited, or if an individual’s outcome depends on the cooperative efforts of participants. When SUTVA holds, the entire population of interest can be described in terms of potential outcomes that contain all of the information needed to ascertain causal effects.

The RCM is a conceptual model, not a practical approach to estimating causal effects because of what Holland (1986) calls the fundamental problem of causal inference— we can never directly estimate \( Y_i(1) - Y_i(0) \) because we can observe an individual after exposure to only one condition. If, however, we are interested in the typical impact that a treatment has on a population, there are a number of ways to sidestep the problem. One is to define the value of interest as the ATE. At a conceptual level, this is defined as \( E[Y_i(1)] - E[Y_i(0)] \), where \( i \) is not equal to \( j \) and exposure to conditions 1 or 0 is assigned at random. At a practical level, the ATE can be calculated as the mean outcome for those exposed to the treatment less the mean outcome for those who were not, again given that participants were assigned to conditions at random. This calculated value is the observed ATE. Because it is based on observed test scores it will
be smaller than the true ATE because of the unreliability of the test. Following Cohen (1988, p. 536), the observed ATE can be inflated to produce an estimate of the true ATE by dividing it by the square root of the reliability of the posttest. When this inflated estimate is expressed as a standardized effect size (denoted by *), it can be calculated as

\[
\text{ATE}^* = \frac{\text{ATE}_{\text{observed}}}{s(y)_{\text{pooled}} \sqrt{r_{yx}}},
\]

where \( s(y)_{\text{pooled}} \) is the pooled standard deviation of the posttest (Lipsey, 1990, p. 78).

It is important to note that average treatment effects are crude measures of a program’s impact; they can be inaccurate if subject-by-treatment interactions result in variable program impacts (Gadbury & Iyer, 2002) and misleading if a program that is demonstrably helpful on average is in fact harmful to specific subgroups (Horowitz, Singer, Makuch, & Viscoli, 1996). Nonetheless, they are the common currency of impact evaluations, the focus of high-level policymakers, and the raw material in most meta-analyses; estimating them well is of critical importance.

**Statistical Models: OLS Regression**

The ATE can be estimated in many ways. I consider estimates that are produced with two ordinary least squares linear regression models, one that utilizes the covariance adjustment approach

\[
y_i = \hat{\alpha}_C + \hat{\beta}_C t_i + \hat{\gamma}_C x_i + \hat{\epsilon}_C i
\]

(Model C)

and the other the gain score approach

\[
y_i - x_i = \hat{\alpha}_G + \hat{\beta}_G t_i + \hat{\epsilon}_G i
\]

(Model G)

In these models, \( x_i, y_i, \) and \( (y_i - x_i) \) refer to the observed pretest, posttest, and gain scores for subject \( i \). The indicator \( t_i \) takes the value of 1 if person \( i \) was assigned to the treatment condition and 0 otherwise. The parameter estimates \( \hat{\alpha}, \hat{\beta}, \hat{\gamma}, \) and \( \hat{\epsilon}_i \) bear subscripts (C or G) that indicate the model in which they appear. The \( \hat{\alpha}s \) estimate the regression intercepts, the \( \hat{\beta}s \) estimate the ATE, \( \hat{\gamma} \) reflects the strength of association between pretest and posttest scores, and the \( \hat{\epsilon}s \) are normally distributed errors with a mean of 0 and a constant standard deviation (Chatterjee, Hadi, & Price, 2006). The parameter of primary interest is \( \beta \), the ATE, and we want to choose between the estimates \( \hat{\beta}_G \) and \( \hat{\beta}_C \) in a way that minimizes bias, maximizes power, and maintains the nominal Type I error rate.
The Gap in Research Design: Intended versus Realized

The debate within the evaluation community concerning experimental methods notwithstanding (Donaldson, 2009), it is generally held that program impacts are estimated most accurately by evaluations that have experimental designs (Rubin, 1974). There are many designs that fall under the heading of experiment, but in recent years randomized trials have been broadly discussed and strongly recommended (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). I limit my investigation to single-level randomized trials in which the units randomly assigned to conditions (e.g., students) are also the units that are measured to ascertain the effectiveness of the program.

In applied settings, there is an unavoidable gap between the ideal experiment that researchers would like to implement and the realized experiment upon which they base their estimates. There are countless ways that realized designs can fall short of what researchers considered ideal or intended to implement, and the resulting gap can have a material effect on a study’s conclusions. For example, attrition can diminish the comparability of treatment and control groups thereby biasing results, typical growth over time can render once difficult tests too easy, and unreliable measures can obscure meaningful program impacts. Often, evaluators do not know the extent or nature of this gap because it cannot be directly observed in the available data. It may be possible, for example, to determine that a measure lacks sufficient reliability to be trusted, but researchers can never know the extent to which any particular configuration of treatment and control groups are truly comparable. Consequently, researchers use their intuition and judgment to make necessary but untestable assumptions regarding the structure of their data. The purpose of this paper is to test a method for choosing between the covariance adjustment and gain score approaches by applying a rule rather than making intuitive assumptions.

Criteria for Choosing Between Approaches Discussed in the Literature

By and large, researchers have suggested two broad criteria that should be used when choosing between approaches—precision and bias. Both have intuitive appeal, yet neither offers a fully satisfactory basis for making the choice when researchers set out to analyze data from flawed experiments.

The Criterion of Precision

Over the past 60 years, there have been repeated claims that gain scores are inherently unreliable and, as a consequence, that the gain score approach may be inappropriate in experimental contexts. Lord (1956, p. 429) argued that because “differences between scores tend to be much more unreliable than the scores themselves” gain scores can produce “absurd results” (p. 421). Cronbach and Furby (1970, p. 68) took the argument a step further, calling on researchers to abandon gain scores altogether:

"Raw change" or "raw gain" scores formed by subtracting pretest scores from posttest scores lead to fallacious conclusions, primarily because such scores are systematically related to any random error of measurement. Although the unsuitability of such scores has long been discussed, they are
still employed, even by some otherwise sophisticated investigators...the authors argue that gain scores are rarely useful, no matter how they may be adjusted or refined.

Similarly, many others have argued that the unreliability of gain scores poses a significant problem for researchers, including McNemar (1958), Manning and Du Bois (1962), Campbell and Stanley (1963), Thorndike (1966), Humphreys and Drasgow (1989), and Bergman (1996). These warnings were taken seriously, to the point that researchers avoided gain scores and some journals established policies that authors must find alternatives to their use (Collins, 1996).

Other researchers countered that assumptions underlying the measure of reliability had led researchers to overstate the problem. As a starting place, they considered the classical test theory definition of reliability, which is

\[ \text{reliability} = \frac{\sigma_{\text{TRUE}}^2}{\sigma_{\text{OBSERVED}}^2} = \frac{\sigma_{\text{TRUE}}^2}{\sigma_{\text{TRUE}}^2 + \sigma_{\text{ERROR}}^2}, \]  

the ratio of the variance of the errorless true scores to the variance of error-prone observed scores (Crocker & Algina, 1986). In the case of gain scores, Equation (2) can be rewritten as

\[ \rho_{(Y-X)} = \frac{\sigma_{\text{TRUE}}^2}{\sigma_{\text{OBSERVED}}^2} = \frac{\sigma_X^2 \rho_X + \sigma_Y^2 \rho_Y - 2 \sigma_X \sigma_Y \rho_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2 \rho_{XY}}, \]  

where \( \rho_{XY} \) is the correlation of the pre and post measures, and \( \rho_{(X-Y)} \), \( \rho_X \), and \( \rho_Y \) are the reliabilities of the gain scores, pretest, and posttest, respectively (Guilford, 1954). Often, researchers make simplifying assumptions when considering the implications of Equation (3), one of which is to assume that the variances of the pretest and posttest are equal. Guilford noted that when this is done, another simplifying assumption, setting \( \rho_X = \rho_Y = \rho_{XY} \), yields a reliability of zero. More generally, Rogosa and Willett (1983), Sharma and Gupta (1986) and Williams and Zimmerman (1996) demonstrated that assuming equal pretest and posttest variances minimizes the reliability of gain scores. The variances need not be equal, and under many realistic circumstances, “an effective experimental treatment which produces changes in the performance of individuals, if measured by an appropriate instrument, can be expected to yield individual gain scores at least as meaningful as other [social science] measures” (Zimmerman & Williams, 1982, p. 154).

Others made the stronger claim that not only is it possible for gain scores to yield sufficiently precise estimates, they are likely to improve precision and in turn increase statistical power. For example, Little and Rubin (2000) explained that when pretest and posttest scores are correlated, using gain scores as outcomes shortens confidence intervals
and improves statistical power (p. 131), and Stuart (2007) stated that under the same condition of correlated tests, “the use of gain scores can yield more efficient estimates” (p. 196). These conclusions rest on the (often implicit) comparison of the gain score approach to a two-sample t-test (equivalent to a regression model with the posttest as the outcome variable, a treatment indicator, and no other covariates). Comparing the gain score approach to the covariate adjustment approach, however, leads to a different conclusion. In the context of a well-implemented experiment, the statistical power afforded by the covariance adjustment approach will always be greater than or equal to that afforded by the gain score approach (Oakes & Feldman, 2001; Porter & Raudenbush, 1987).

To see how this last point applies to experimental designs, Model G can be rewritten as

\[ Y_i = \alpha_{Gi} + \beta_{Gi} T_i + X_i + \varepsilon_{Gi}. \]  

(Model G alt.)

The coefficient of \( X_i \) is fixed to the constant 1 rather than being estimated as it is in Model C. Thus, Model C will always minimize the residual variance in an OLS regression and Model G will not, unless purely by chance the regression coefficient for \( X_i \) equals 1 (in which case the models are equivalent). The resultant loss of power led some to question the usefulness of the gain score approach (Feldt, 1957; Reichardt, 1979), while others, for example Allison (1990), countered that in quasi-experimental contexts it is reasonable to assume the regression coefficient is 1 and the assumption is unlikely to impact results if it turns out to be wrong (see Oakes & Feldman, 2001, pp. 11-12).

The arguments above suggest that neither approach affords greater precision in all circumstances. Starting from this premise, Williams and Zimmerman (1982) suggested that researchers should choose the approach with the more reliable outcome measure (the outcome measure associated with the covariance adjustment approach is the residualized gain score—the posttest after controlling for variation explained by the pretest—not the unadjusted posttest). Like Guilford, they pointed out how the relative magnitudes of the pretest and posttest variances affected reliability, and from this they concluded that the most salient relationship in Equation (3) is the ratio of the pretest and the posttest standard deviations as compared to the pretest-posttest correlation. This led them to construct the following rule:

\[
\begin{align*}
\rho_{XY} &> \frac{\sigma_X}{\sigma_Y} & \text{use gain score approach} \\
\rho_{XY} &< \frac{\sigma_X}{\sigma_Y} & \text{use covariance adjustment approach} \\
\rho_{XY} &= \frac{\sigma_X}{\sigma_Y} & \text{reliabilities are equal; use either approach.} 
\end{align*}
\]  

(4)
While this rule has the virtue of being simple, Zumbo (1999) noted that its simplicity was achieved by ignoring possible data structures that Williams and Zimmerman believed were unlikely to occur in practice. In an effort to be more comprehensive, Zumbo expanded Williams and Zimmerman’s single rule to five. To help researchers select the appropriate rule, he created a flowchart (presented in a slightly modified form in Figure 1). As indicated in the figure, researchers can navigate the flow chart by comparing the variances, reliabilities, and correlations of pretest and posttest scores. (Note that Figure 1 includes two complicated lower bounds, $\psi$ and $\psi'$, that do not enter into the topic under consideration here and are therefore not discussed.)

![Flowchart](image-url)

Figure 1. A flowchart to help decide whether to use the covariance adjustment or gain score approach originally presented in Zumbo (1999, Figure 2 (modified)).

As commonsensical as these rules may be, the relationship between the reliability of gain scores and statistical power is not as straightforward as the rules might suggest. Overall and Woodward (1975) noted what they called the reliability-power paradox—all else being equal, the power of a t-test is maximized when the reliability of the gain scores
Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches

is zero. Their conclusion rests on the fact that the reliability as defined in Equation (3) decreases as true score variance decreases. This can occur, for example, when the conditions that Guilford noted are realized, and it would be a serious problem if the aim of the research were to describe individual differences in gains. When the purpose is to estimate a causal effect, however, true score variance is a nuisance because it obscures treatment effects with unrelated variance. As true score variance decreases, the study sample becomes more homogenous (i.e., its variance grows smaller), which increases statistical power (Lipsey, 1990, pp. 77-78).

Equation (3) reflects another paradox—holding all else equal, the reliability of gain scores, \( \rho_{xy} \), increases as the pretest-posttest correlation, \( \rho_{xy} \), decreases. Bereiter (1962) dubbed this the unreliability-invalidity dilemma because psychometricians typically strive to design measures with high test-retest reliability because it provides evidence that the construct and its measure are stable over time, yet this desirable property is associated with lower gain score reliability. Likewise, experimental researchers are also caught in this paradox because they strive to maximize pretest-posttest correlations in order to maximize the power of the statistical models they use to detect causal effects (Bloom, Richburg-Hayes, & Black, 2007). If one set out to obtain a lower value for \( \rho_{xy} \) in the hope of increasing gain score reliability, one might be using a measure that does not reflect the same construct over time and undermines the power of statistical tests. Bohrnstedt (1969) concluded that although the dilemma was overstated (low test-retest reliability does not necessarily imply the measurement of different constructs), it raised serious questions about the reliability of gain scores that further confused the issue of how they should be used in experimental settings.

The reliability-power paradox and the unreliability-invalidity dilemma contradict the direct relationships between statistical power and two properties of a measure—reliability and pre-post correlations—that researchers have traditionally depended on when designing experiments. Not surprisingly, some researchers pushed back against the notion that this direct relationship does not hold (see for example, Humphreys, 1996). However, the finding that “in research concerned with differences between experimental treatment groups, the loss in reliability due to calculation of difference scores is not a valid concern,” (Overall & Woodward, 1975, p. 85) has stood the test of time (Allison, 1990). That is not to say that the reliabilities of the pretest and posttest scores do not matter; both increase the reliability of gain scores as well as the power of the gain score and covariance approaches (Humphreys & Drasgow, 1989). But it does undermine the premise that one should select the approach with the more reliable outcome measure.

The Criterion of Bias

A biased estimate is one that on average does not equal the true parameter value being estimated (Hogg & Craig, 1995, p. 263). When an experiment is well implemented, both the covariance adjustment and gain score approaches produce unbiased ATE estimates (Maxwell & Howard, 1981), but when an experiment is degraded, the two approaches may yield estimates that are differentially biased, depending on the true data generating process. With this in mind, Allison (1990)
examined how researchers could choose between approaches in order to minimize bias when conducting non-experimental studies.

<table>
<thead>
<tr>
<th>Case</th>
<th>$ATE = 0$</th>
<th>$ATE &gt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Pre</td>
<td>Post</td>
</tr>
<tr>
<td>Intact Experiment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonequivalence at Baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differential Growth</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. The three cases discussed by Allison (1990)—an intact experiment, nonequivalence at baseline, and differential growth. Treatment (+) and control (O) means are plotted for programs with no effect ($ATE=0$) and a positive effect ($ATE>0$).

He considered three data structures that I have labeled Cases A, B, and C in Figure 2. Case A occurs when random assignment, as it plays out in real-world circumstances, produces the intended long-run equivalence of groups. Thus, the treatment and control groups can be expected to have the same average pretest scores, and posttest scores will tend to differ by the true ATE (attenuated by the unreliability of the measure). Growth over time may occur, but the expected growth in both groups is equal and therefore does not affect ATE estimates. Case B occurs when the true data generating mechanism tends to yield groups that have different means on the pretest. When the true ATE is zero, the mean difference between groups on the posttest will tend to be equal to the mean difference on the pretest. Otherwise, the mean treatment-control difference on the posttest will tend to be larger or smaller than the true ATE, depending on whether the mean difference on the pretest was positive or negative. This can occur if
the randomization procedure is flawed, if there is differential attrition across groups, or the pretest is administered after participants are aware of the condition to which they have been assigned. Under Case B, the two approaches yield different ATE estimates on average. Allison noted that in this case there is a strong consensus that the gain score approach will yield ATE estimates that are less biased than the covariance adjustment approach.

Case B has received a great deal of attention in the literature considering the choice between gain score and covariance adjustment approaches, in part because of its strong association with Lord’s paradox (Lord, 1967). Nonetheless, although Case B offers a plausible data structure for non-experimental or degraded experimental studies, it is not the only plausible structure. For example, Case C occurs when, in the absence of an effective treatment, participants in the two groups tend to grow (or decline) at different rates. Allison noted that neither the gain score nor the covariance adjustment approach properly adjusts for differential growth in this instance because growth is completely confounded with treatment effects.

Because there are many ways that the structure of non-experimental data may undermine estimates of treatment effects, Allison concluded that neither the gain score approach nor the covariance adjustment approach would be superior in every situation. He suggested researchers choose an approach based on a subjective determination of whether the implications of an approach fit with what the researcher intuitively expected for a given phenomenon. To illustrate this point, he presented results from a quasi-experiment that corresponds to Case B in Figure 2. In one instance, the data produced a plot that appeared similar to the one associated with an ATE of zero. The covariance adjustment approach yielded a non-zero estimate of the causal effect while the gain score approach yielded an estimate of zero. Because inspection of the data should lead us to expect that the program had no effect, the gain score approach should be preferred in this case. In the other instance, the data produced a plot similar to the one associated with a non-zero ATE. Here, the covariance adjustment approach yielded an estimate of zero and the gain score approach a non-zero estimate. Again, because inspection of the data should lead us to expect that the program had an effect, the gain score approach should be preferred in this case.

Similarly, Wainer (1991) also argued that neither approach was always superior and the appropriateness of an approach depended on the untestable assumptions made by the researcher. He demonstrated this by first rewriting the ATE estimate produced by the gain score approach as

\[ \hat{\beta}_g = \frac{E(Y'_1 - E(X_1)) - E(Y'_0 - E(X_0))}{E(Y'_1) - E(Y'_0)} \]

In this form, the difference in the expected values of the pretest scores can more easily be identified as a bias term that is subtracted from the difference in expected values of the posttest scores. On the other hand, the estimate produced by the covariance adjustment approach is
Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches

\[ \hat{\beta}_c = \hat{\alpha}_{c1} - \hat{\alpha}_{c0}, \]  

(6)

the difference in the estimated intercepts for the regression lines for the treatment and control groups.

Figure 3 illustrates these two estimates graphically for situations that conform to Cases A and B. When Case A holds, \( E(X_1) = E(X_0) \), so there is no bias as represented in Equation (5) and the gain score estimate reduces to the vertical distance \( E(Y_1) - E(Y_0) \).

This is equivalent to the vertical distance \( \alpha_{c1} - \alpha_{c0} \), thus the two approaches produce the same ATE estimates over the long run. When Case B holds, \( E(X_1) \neq E(X_0) \) and the bias term in Equation (5) is not zero. Consequently, the ATE estimates produced by the two approaches will tend to differ over the long run, except when the slopes of the parallel regression lines are 1. In this unlikely event, as the horizontal distance \( E(X_1) - E(X_0) \) increases, the vertical distance \( E(Y_1) - E(Y_0) \) increases by the same amount, thereby offsetting each other and reducing Equation (5) to Equation (6).

![Figure 3](image)

Figure 3. A graphical representation of estimates produced by the covariance adjustment and gain score approaches for Cases A and B.

Like Allison, Wainer argued that researchers should choose an approach after inspecting their data and forming a belief about counterfactuals. If in the absence of the treatment it is reasonable to assume that there will be no change over time, Wainer suggested that the gain score approach should be preferred. If change over time is assumed, the covariance adjustment approach should be preferred. Although this may be a useful rule for non-experimental studies that allow researchers to observe multiple pre-
intervention measures and incorporate them in their analysis (as was the case with Wainer’s example), it is less useful in experimental settings (potentially flawed as they may be) where researchers can observe only pre- and post-intervention measures. Consider Case B as illustrated in Figure 3. The assumption of constant growth over time implies that both lines would be shifted up or down by an equal amount, which has no effect on the ATE estimates of either approach.

Maris (1998) took a different tack when considering how to choose between approaches in non-experimental settings. He investigated how knowledge of the selection mechanism affected ATE estimates and concluded that if the assignment mechanism was on the basis of pre-intervention abilities, only the covariance adjustment approach was unbiased. The reason for this is that the pretest functions as a propensity score (Rosenbaum & Rubin, 1983) because it is associated with the probability of being assigned to the treatment group. If the assignment mechanism is not known, Maris argued that there is no basis for selecting one approach over another.

Table 1

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Slopes parallel</th>
<th>Slopes not parallel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Covariance</td>
<td>Gain</td>
</tr>
<tr>
<td></td>
<td>Adjustment</td>
<td>Score</td>
</tr>
<tr>
<td>Random</td>
<td>Not Considered</td>
<td>Not Considered</td>
</tr>
<tr>
<td>Baseline score determines group assignment</td>
<td>Unbiased</td>
<td>Biased</td>
</tr>
<tr>
<td>Baseline score is associated with but does not determine group assignment</td>
<td>Biased</td>
<td>Unbiased</td>
</tr>
</tbody>
</table>

Wright (2006) explored the effect of assignment mechanisms in more detail, constructing simulations of five non-experimental conditions (see Table 1). Assignment to conditions was modeled as random, determined by baseline scores (considered by Maris), or associated with but not determined by baseline scores. In addition, the treatment and control group regression lines were either parallel (as they appear in Figure 3) or not. When assignment was random, both approaches yielded unbiased estimates when slopes were not parallel (he did not consider the paradigmatic case of random assignment and parallel slopes). When the baseline score determined group assignment, the covariance adjustment approach alone provided unbiased estimates regardless of the nature of the slopes. And when the baseline score was associated with but did not determine group assignment (a mechanism that can yield Case C), the nature
of the slopes was relevant; the gain score approach produced unbiased estimates when slopes were parallel, but neither approach yielded unbiased estimates otherwise.

Discussions of Past Research

Several conclusions can be drawn from the research described above. First, the reliability of the outcome measure does not appear to be a useful criterion for selecting between approaches in experimental contexts. Perhaps the most compelling reason to consider the reliability of outcome measures in experiments is to ensure that there is sufficient statistical power. Because the relationship between the reliability of gain scores and power is not straightforward, it would seem better to judge the utility of the approaches by the power they actually afford rather than a weak indicator of it.

Second, past research has for the most part evaluated approaches using one criterion at a time. It would be desirable to consider multiple criteria, in particular bias, power, and the Type I error rate. Greater sensitivity to detect an effect may not be desirable if it results from estimates that are biased upward. And decreasing the bias of an estimate may not be desirable if doing so compromises the Type I error rate of statistical tests.

Third, the choice of approaches depends on the underlying data structure. Allison (1990), Maris (1998), and Wright (2006) make it clear that the assignment mechanism (random or probabilistic), associations with the assigned condition (different pretest means and pretest-posttest slopes across treatment and control groups), and counterfactual conditions (differential growth in the absence of the treatment) affect the results produced by the two approaches. In experimental contexts, even when the experiment has been degraded, the assignment mechanism, per se, cannot affect the ATE estimate because it was random. However, in degraded experiments conditions may be associated with differential means (Case B) and differential growth (Case C), and in both cases slopes may or may not be parallel.

Fourth, relying on the judgment of the researcher to properly assess the underlying data structure is problematic. Past research suggests that researchers should apply a mix of approaches depending on the underlying data structure. Not all of the relevant structure may be observed, however, and some observable features may be the result of sampling error rather than systematic structure. Moreover, researchers may disagree about what they see and how they should respond to it. An objective procedure for judging structure and taking action would be more desirable.

A Mixed Approach

Rather than rely solely on the intuition of researchers who visually inspect their data, I propose an objective criterion—comparing an estimate of the true difference between treatment and control groups at the time of the pretest, $\hat{\lambda}$, to a predetermined cutoff, $c$. If $\hat{\lambda}$ is greater than $c$, the gain score approach should be used, otherwise the covariance adjustment approach should be used. I refer to this as the mixed approach because it uses a mixture of the gain score and covariance approaches in all situations.

The estimate $\hat{\lambda}$ can be calculated in three steps. First, the absolute value of the mean treatment-control difference on the pretest is calculated. This represents the
magnitude of the bias expressed in Equation (5) assuming Case B holds. Second, the
difference is standardized using the pooled standard deviation of the posttest. Third, the
standardized value is divided by the square root of the reliability of the outcome measure
to disattenuate the difference. This puts \( \hat{\lambda} \) in the same units as ATE* estimate (the
standardized ATE in Equation (1)) that is the object of the experiment. This three-step
procedure can be summarized as

\[
\hat{\lambda} = \frac{|\bar{x}_1 - \bar{x}_0|}{s(y)} \cdot \frac{1}{\sqrt{r_{yy}}}. \tag{7}
\]

Although \( \hat{\lambda} \) can be estimated, the cutoff, \( c \), must be set to a specific value. The suggested
value, based on the results of the simulations that follow, is 0.20. The mixed approach is
intended to help balance three objectives—minimizing bias, maximizing statistical
power, and maintaining the nominal Type I error rate, and it is evaluated on how closely
it achieves these ideal values.

Simulations of Flawed Experiments
In order to benchmark the performance of the three approaches when an
experiment is flawed, I simulated multiple datasets for various “data scenarios” that entail
a number of assumptions about the structure of the data. I then estimated the ATE for
multiple iterations of each dataset using the two regression models described above
(Models C and G) and compared the performance of the covariance adjustment, gain
score, and mixed approaches using three criteria (bias, power, and Type I error rate).
Simulations were executed in S-Plus (TIBCO Software Inc., 2009) using code developed
by the author (see Appendix A). The data scenarios and data generation procedures are
described below.

Data Scenarios
Data scenarios are labeled A1-A4 and B1-B4 (see Figure 4). Each scenario
corresponds to either Case A or Case B (designated by A or B); Case C it is not
considered because it is known that neither approach performs well in this case. When
either Case A or B holds, four conditions might influence the relative performance of the
gain score and covariance adjustment approaches (designated by the numbers 1-4). In
Scenarios A1 and B1 (Equal Variance), the standard deviations of the pretest and posttest
ability distributions are equal, which minimizes the reliability of the gain scores. In
Scenarios A2 and B2 (Increasing Variance), the standard deviation of the posttest ability
distribution is greater than that of the pretest ability distribution. This increases the
reliability of the gain scores. In Scenarios A3 and B3 (Extreme Growth), there is
substantial growth over time that is common to the treatment and control groups. In
Scenarios A4 and B4, the impact of the treatment is a function of pretest abilities; thus,
Models C and G are incorrectly specified because they do not include an appropriate
interaction term. Data for these eight data scenarios were generated by manipulating the
parameters listed in Figure 4 in a two-step process that is described below.
### Figure 4. The relationship of cases, data scenarios, and parameter values used to simulate data. Treatment (+) and control (O) means are plotted for \( \beta=0 \) and \( \beta=0.50 \).

The true abilities of participants were generated by first drawing *unadjusted* abilities form a bivariate normal distribution. These values are denoted \( \theta_{im}^* \), where \( m \) indexes the time of the pretest (1) and posttest (2). The bivariate normal distribution was specified such that \( \theta_{1m}^* \) was distributed \( N(0,1) \); \( \theta_{2m}^* \) was distributed \( N(0, \tau) \), where \( \tau \) could take the value of 1 or \( \sqrt{2} \); and \( \rho(\theta_{1m}^*, \theta_{2m}^*) \) was set to five different values—0, .25, .50, .75, and .99—in order to investigate how the strength of this correlation affected ATE estimates. Throughout this paper, the other parameters in Figure 4—\( \phi, \lambda, \beta, \) and \( \omega \)—are expressed as disattenuated standardized effects, and they are standardized using the pooled standard deviation of the posttest. In order to convert the values in Figure 4 to the units of the posttest when \( \tau \neq 1 \), the parameter values were multiplied by \( \tau \), the standard deviation of the unadjusted posttest ability distribution, as described in the equations that follow.

<table>
<thead>
<tr>
<th>Case</th>
<th>( \beta=0 )</th>
<th>( \beta=0.50 )</th>
<th>Data Scenario</th>
<th>( \tau )</th>
<th>( \phi )</th>
<th>( \lambda )</th>
<th>( \omega )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Intact</td>
<td></td>
<td></td>
<td>A1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Experiment</td>
<td></td>
<td></td>
<td>A2</td>
<td>( \sqrt{2} )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A3</td>
<td>( \sqrt{2} )</td>
<td>2.5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A4</td>
<td>( \sqrt{2} )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>Nonequivalence</td>
<td></td>
<td></td>
<td>B1</td>
<td>1</td>
<td>0</td>
<td>-0.5</td>
</tr>
<tr>
<td></td>
<td>at Baseline</td>
<td></td>
<td></td>
<td>B2</td>
<td>( \sqrt{2} )</td>
<td>0</td>
<td>-0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B3</td>
<td>( \sqrt{2} )</td>
<td>2.5-0.5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B4</td>
<td>( \sqrt{2} )</td>
<td>0</td>
<td>-0.5</td>
</tr>
<tr>
<td>C</td>
<td>Differential</td>
<td></td>
<td></td>
<td>Not Considered</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Growth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Step 1: Generating True Abilities

The unadjusted abilities, $\theta^*_m$, were adjusted to create $\theta_m$, the realized abilities at time $m$. The realized abilities at the time of the pretest were calculated as

$$\theta_{i1} = \theta^*_i + \lambda t_i,$$

where $t$ is a randomly generated binary variable that indicates the group to which participant $i$ was assigned, $\lambda$ is used to create Case B by imposing a difference in means between the treatment and control groups at the time of the pretest, and $\tau$ converts $\lambda$ to the units of the posttest. Note that when $\lambda \neq 0$, the variance of $\theta_{i1}$ will tend to be slightly greater than 1.

The realized ability at the time of the posttest can be thought of as the random assignment of one of two potential outcomes. The outcome after exposure to the control condition was calculated as

$$Y_i(0) = \theta^*_{i2} + \tau \phi,$$

and the outcome after exposure to the treatment condition was calculated as

$$Y_i(1) = \theta^*_{i2} + \tau \left[ \phi + \lambda + \beta + \omega(\theta_{i1} - \bar{\theta}_i) \right].$$

Here, $\phi$ represents growth that is common to the treatment and control groups; $\lambda$ is the persistence of the baseline difference; $\beta$ is the true ATE, the parameter that we wish to estimate; and $\omega$ determines the strength of a treatment-by-pretest-ability interaction. Again, the parameter values were multiplied by $\tau$ in order to convert them to the units of the posttest. A potential outcome is assigned to participants using a randomly generated treatment indicator, $t$, so the realized outcome can be calculated in one step as

$$\theta_{i2} = \theta^*_{i2} + \tau \left( \phi + t_i \left[ \lambda + \beta + \omega(\theta_{i1} - \bar{\theta}_i) \right] \right).$$

Note that if some combination of $\lambda$, $\beta$, and $\omega$ differ from zero, the variance of $\theta_{i2}$ may differ from $\tau$ and $\rho(\theta_{i1}, \theta_{i2})$ may differ from $\rho(\theta^*_i, \theta^*_2)$. Thus, $\tau$ and $\rho(\theta^*_i, \theta^*_2)$ can be thought of as the expected standard deviation of the posttest ability distribution and pre-post correlation of true abilities when random assignment is fully intact and no treatment is administered, respectively. For each of the five values of $\rho(\theta^*_i, \theta^*_2)$, 3000 datasets were generated, each of which provided data for 60 subjects with no missing data.
Step 2: Generating Test Scores

Pretest and posttest sum scores were generated by simulating the responses to 60 dichotomously scored test items (correct=1, wrong=0) that conformed to a Rasch model (Rasch, 1960/1980). In this model, the probability that person $i$ responds correctly to item $j$ at time $m$ is a function of $\theta_{im}$ and the difficulty of item $j$, denoted $\delta_j$, such that

$$P(X_{im} = 1) = \frac{\exp(\theta_{im} - \delta_j)}{1 + \exp(\theta_{im} - \delta_j)}.$$ \hspace{1cm} (12)

The 60 item difficulty parameters used in these simulations were created by evenly dividing the interval between -2 and 2 logits, yielding a sum score reliability (Cronbach’s alpha) of .91 when the standard deviation of the underlying ability distribution is 1 (and there are not ceiling or floor effects). All else being equal, the greater the variance of the ability distribution (i.e., the true scores), the greater the reliability of the observed scores (see Equation (2)). Thus, when $\tau = \sqrt{2}$, the reliability of the posttest exceeds that of the pretest.

![Figure 5](image-url)

Figure 5. The many-to-many mapping of true abilities to sum scores, and the one-to-one mapping of sum scores to Rasch ability estimates.
The response to item \( j \) for person \( i \) at time \( t \) was drawn at random from a Bernoulli distribution using \( P(X_{im} = 1) \) as the probability of success. These responses were then summed to produce observed pretest and posttest scores for each person. As illustrated in Figure 5, this produces a many-to-many mapping from true abilities to sum scores—two individuals with the same true ability can give rise to different sum scores, and the same sum score could have been produced by individuals with different true abilities. When a test is reliable, individuals with the same true ability tend to give rise to the same or very similar sum scores, and one sum score tends to map to a very narrow range of true abilities that has little overlap with other ranges.

One could take the simulated item responses and then estimate the underlying abilities by fitting a Rasch model using one of several possible maximum likelihood procedures (Embretson & Reise, 2000). The choice to use sum scores rather than Rasch estimates was made for two reasons. First, when there are no missing data, as with these simulations, there is a one-to-one mapping of sum scores to Rasch estimates (see Figure 5) because sum scores are a sufficient statistic for the Rasch estimates (Molennar, 1995). That is, Rasch estimates contain no additional information beyond what is contained in the sum scores when all data are available, although the shape of the distributions may differ slightly (the Rasch estimates tend to have longer tails). Second, sum scores are widely used by practicing evaluators. They are more familiar, simpler to estimate using the count-the-number-correct conventional scoring rule (Crocker & Algina, 1986, p. 399), and easier to communicate to non-specialists (Sijtsma & Hemker, 2000). For evaluators, who often do not have extensive quantitative training, these may be compelling properties.

It is important to include Step 2 because the practical relationship between true abilities and observed scores has not been well explored in previous work addressing the choice between gain score and covariance adjustment approaches. A principal concern is how well the item difficulties cover the range of true abilities for the sample of participants being assessed. In Figure 6, distribution B illustrates the ideal case. The average difficulty of the items aligns with the middle of the range of true abilities, so a participant in the middle of the range of abilities has a 50% chance of correctly answering an item that is in the middle of the range of item difficulties. Distributions A and C, however, are not well aligned. The test is too difficult for the sample described by distribution A, and this results in a floor effect that shifts the observed distribution’s mean to the right of its median, reduces its variance, and reduces its reliability. On the other hand, the test is too easy for the sample described by distribution C, and this results in a ceiling effect that has all of the same consequences except that it shifts the mean of the observed score distribution to the left. By including Step 2, it allows us to see how simple changes in the mean and variance of the true ability distributions can trigger complicated distortions in the observed distribution when the same test is administered twice, and how in turn this can impact ATE estimates and our preference for the approach we use to estimate it.
Estimating the ATE

The sum scores generated in Step 2 were used to fit Models C and G. For both models, ATE estimates were standardized and disattenuated as described in Equation (1) and these estimates and associated p-values were stored. For the mixed approach, \( \hat{\lambda} \) was estimated as described in Equation (7) and compared to values of \( c \) ranging from 0.05 to 0.35 in increments of 0.05. For \( \hat{\lambda} > c \), estimated values produced by the gain score approach were used, otherwise estimated values produced by the covariance adjustment approach were used. ATE estimates and findings of statistical significance (quantified as 0 or 1) for the covariance adjustment, gain score, and mixed approaches were averaged to empirically estimate (a) the degree of bias when the true ATE was 0.50 standard deviations, (b) the power to detect an effect of that size, and (c) the Type I error rate when the ATE was 0.
Results

Results for Scenario A1: Well-Implemented Experiments where $\sigma_x = \sigma_Y$

As discussed above, both the covariance adjustment and gain score approaches should provide unbiased estimates and realize the nominal Type I error rate when experiments are well implemented. However, the covariance adjustment approach will tend to be more powerful and should therefore be preferred in Scenario A1. When the mixed approach is used, it will sometimes select the less powerful gain score approach, and as indicated in Table 2 this resulted in a small penalty—power was slightly lower and the Type I error rate was slightly elevated (no penalty was paid in terms of bias). The magnitude of the penalty depended on the size of the cutoff value, $c$, and the pre-post correlation of the true abilities. As the cut off decreased, power decreased and the Type I error rate approached the nominal rate, so in choosing a value for $c$ there can be tension between maximizing power and maintaining the nominal error rate. The most advantageous choice for the cutoff is not clearly indicated in this table, but from inspection of all results 0.20 appears to work well and produce acceptable tradeoffs.

Table 2
Simulation Results for Scenario A1: Bias, Power, and Type I Error Rates for the Covariance Adjustment, Gain Score, and Mixed Approaches

<table>
<thead>
<tr>
<th>$\rho(\theta_{i1}, \theta_{i2})$</th>
<th>0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta=0.50$</td>
<td>$\beta=0$</td>
<td>$\beta=0.50$</td>
<td>$\beta=0$</td>
<td>$\beta=0.50$</td>
</tr>
<tr>
<td><strong>Covariance Adjustment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model C</td>
<td>0.00</td>
<td>0.44</td>
<td>0.05</td>
<td>0.01</td>
<td>0.45</td>
</tr>
<tr>
<td>Mixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = 0.35$</td>
<td>0.00</td>
<td>0.42</td>
<td>0.07</td>
<td>0.00</td>
<td>0.43</td>
</tr>
<tr>
<td>$c = 0.30$</td>
<td>0.00</td>
<td>0.42</td>
<td>0.07</td>
<td>0.00</td>
<td>0.42</td>
</tr>
<tr>
<td>$c = 0.25$</td>
<td>0.00</td>
<td>0.40</td>
<td>0.07</td>
<td>0.00</td>
<td>0.41</td>
</tr>
<tr>
<td>$c = 0.20$</td>
<td>0.00</td>
<td>0.37</td>
<td>0.07</td>
<td>0.00</td>
<td>0.39</td>
</tr>
<tr>
<td>$c = 0.15$</td>
<td>0.00</td>
<td>0.35</td>
<td>0.06</td>
<td>0.00</td>
<td>0.37</td>
</tr>
<tr>
<td>$c = 0.10$</td>
<td>0.00</td>
<td>0.31</td>
<td>0.06</td>
<td>0.00</td>
<td>0.35</td>
</tr>
<tr>
<td>$c = 0.05$</td>
<td>0.00</td>
<td>0.28</td>
<td>0.05</td>
<td>0.00</td>
<td>0.33</td>
</tr>
<tr>
<td><strong>Gain Score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model G</td>
<td>0.00</td>
<td>0.24</td>
<td>0.05</td>
<td>0.00</td>
<td>0.30</td>
</tr>
</tbody>
</table>

As the pre-post correlation increased, the discrepancies in power and error rate between Models C and G decreased. By using a cutoff of 0.20 when the pre-post correlation was 0, power was .37 using the mixed approach versus .44 using the
covariance adjustment approach, and the Type I error rate was .07 versus the nominal rate achieved under Model C. Using the same cutoff when the correlation was .75 yielded power and error rates for the mixed approach of .64 and .06, respectively, versus .67 and .05 for the covariance adjustment approach. Thus, with a reasonably strong pre-post correlation, which is often found in practice, the penalty for using the mixed approach for data conforming to Scenario A1 appears to be negligible.

**Results for Scenario A2: Well-Implemented Experiments where \( \sigma_x < \sigma_y \)**

As with Scenario A1, the covariance adjustment approach was more or equally powerful than the gain score approach, and for both approaches the estimates were unbiased and the Type I error rates reflected the nominal rate. Thus, for all levels of the pre-post correlation the covariance adjustment approach was again preferred. Using a cutoff of 0.20 when pre-post correlation was 0 resulted in power for the mixed approach versus the covariance adjustment approach of .42 versus .45, a trivial difference. The Type I error rate was .06 versus the nominal rate of .05, which again is a very small difference. When the pre-post correlation was .75, there was no difference in power between the two models (.73 in both cases) and Type I error rate for the mixed approach achieved the nominal level.

**Table 3**

*Simulation Results for Scenario A2: Bias, Power, and Type I Error Rates for the Covariance Adjustment, Gain Score, and Mixed Approaches*

<table>
<thead>
<tr>
<th>( \rho(\theta_i^n, \theta_i^o) )</th>
<th>0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta=0.50 )</td>
<td>Bias</td>
<td>Power</td>
<td>Type I Error</td>
<td>Bias</td>
<td>Power</td>
</tr>
<tr>
<td>Covariance Adjustment</td>
<td>Model C</td>
<td>0.01</td>
<td>0.45</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>( c = 0.35 )</td>
<td>0.01</td>
<td>0.45</td>
<td>0.05</td>
<td>0.00</td>
<td>0.45</td>
</tr>
<tr>
<td>( c = 0.30 )</td>
<td>0.01</td>
<td>0.44</td>
<td>0.06</td>
<td>0.00</td>
<td>0.45</td>
</tr>
<tr>
<td>( c = 0.25 )</td>
<td>0.01</td>
<td>0.43</td>
<td>0.06</td>
<td>0.00</td>
<td>0.45</td>
</tr>
<tr>
<td>( c = 0.20 )</td>
<td>0.01</td>
<td>0.42</td>
<td>0.06</td>
<td>0.00</td>
<td>0.43</td>
</tr>
<tr>
<td>( c = 0.15 )</td>
<td>0.01</td>
<td>0.40</td>
<td>0.06</td>
<td>0.00</td>
<td>0.42</td>
</tr>
<tr>
<td>( c = 0.10 )</td>
<td>0.01</td>
<td>0.37</td>
<td>0.05</td>
<td>0.00</td>
<td>0.41</td>
</tr>
<tr>
<td>( c = 0.05 )</td>
<td>0.01</td>
<td>0.34</td>
<td>0.05</td>
<td>0.00</td>
<td>0.39</td>
</tr>
</tbody>
</table>

**Gain Score**

| Model G | 0.01 | 0.31 | 0.04 | 0.00 | 0.37 | 0.05 | 0.00 | 0.49 | 0.05 | 0.01 | 0.73 | 0.05 | 0.00 | 0.99 | 0.05 |

One reason for the improved performance of the mixed approach is that the discrepancy in power between the covariance and the gain score approaches was smaller in Scenario A2 as compared to that in Scenario A1. This can be explained by the
increase in the posttest variance and the resulting increase in the reliability of the posttest, both of which improved the reliability of the gain scores (see Equation (3)). This in turn increased the power of the gain score approach, reducing the penalty in terms of power when the mixed approach chose Model G over Model C.

**Results for Scenario A3: Well-Implemented Experiments with Extreme Growth**

For this scenario, the covariance adjustment approach and the gain score approach provided ATE estimates that were slightly biased by equal amounts (from -0.03 to -0.05 standard deviations). One might consider this to be negligibly small, but it was associated with a substantial reduction in power as compared to Scenario A2 (between 5 and 18 percentage points). The Type I error rate did not differ from the nominal rate. Neither approach provides ideal results, but given that the bias was small the covariance adjustment approach is still preferable because it yielded more powerful statistical tests.

**Table 4**

*Simulation Results for Scenario A3: Bias, Power, and Type I Error Rates for the Covariance Adjustment, Gain Score, and Mixed Approaches*

<table>
<thead>
<tr>
<th>β = 0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>C = 0.35</td>
<td>-0.05</td>
<td>0.34</td>
<td>0.08</td>
<td>-0.03</td>
</tr>
<tr>
<td>C = 0.30</td>
<td>-0.05</td>
<td>0.33</td>
<td>0.08</td>
<td>-0.03</td>
</tr>
<tr>
<td>C = 0.25</td>
<td>-0.05</td>
<td>0.32</td>
<td>0.07</td>
<td>-0.03</td>
</tr>
<tr>
<td>C = 0.20</td>
<td>-0.05</td>
<td>0.29</td>
<td>0.07</td>
<td>-0.03</td>
</tr>
<tr>
<td>C = 0.15</td>
<td>-0.05</td>
<td>0.27</td>
<td>0.07</td>
<td>-0.03</td>
</tr>
<tr>
<td>C = 0.10</td>
<td>-0.05</td>
<td>0.24</td>
<td>0.06</td>
<td>-0.03</td>
</tr>
<tr>
<td>C = 0.05</td>
<td>-0.05</td>
<td>0.20</td>
<td>0.06</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

The reason for the bias is illustrated in Figure 7. The pretest and posttest ability distributions were separated by $\phi$, the average growth (in the absence of the treatment) between the time of the pretest and posttest. Growth was so great (2.5 standard deviations) as to render the posttest too easy. The resultant ceiling effect caused three things to happen—the posttest mean was shifted to the left (denoted by *), the pooled standard deviation of the posttest was decreased, and the reliability of the posttest was decreased. When the true ATE was zero, the control group and treatment group distributions were shifted to the left by an equal amount, $a$ units. Thus the ATE estimate
was zero (unbiased) and the nominal Type I error rate achieved. When the true ATE was greater than zero, the mean of the posttest was shifted $b$ units to the left, where $b > a$. This diminished the ATE estimate in the original units, helping to reduce statistical power. Because the results are reported as a disattenuated standardized effect sizes as described in Equation (1), the smaller pooled standard deviation and decreased reliability inflated this underestimate. These offsetting effects resulted, in this particular set of circumstances, in a slightly smaller estimated ATE and an associated reduction in power.

It is worth noting that the mixed approach was no more biased than either the covariance adjustment or gain score approaches, and the penalty for using the mixed approach in terms of the Type I error rate was as small as in previous scenarios. The loss of power was greater than it was in Scenario A2, however. When the cutoff was set to 0.20 and the pre-post correlation was .75, the mixed approach yielded power of .46 versus .56 for the covariance adjustment approach, which is a material difference.

![Figure 7: The impact of a posttest ceiling effect on Scenario A3.](image)
Results for Scenario A4: Well-Implemented Experiments with Differential Impacts

In this scenario, the magnitude of the treatment effect varied by person as a function of their pre-intervention ability. This differential impact was present when the true ATE was both 0.50 and 0 standard deviations. Thus, Model C and Model G were always the “wrong” models because they did not include interaction effects. The range of the treatment effects in this scenario was greater than ±2.5 standard deviations, which is quite large. Consequently, when the true ATE was zero, the program did great harm and good, and when the ATE was larger than zero the program did good on average but at the cost of many being substantially harmed.

Table 5

Simulation Results for Scenario A4: Bias, Power, and Type I Error Rates for the Covariance Adjustment, Gain Score, and Mixed Approaches

<table>
<thead>
<tr>
<th>β = 0.50</th>
<th>β = 0</th>
<th>β = 0.50</th>
<th>β = 0</th>
<th>β = 0.50</th>
<th>β = 0</th>
<th>β = 0.50</th>
<th>β = 0</th>
<th>β = 0.50</th>
<th>β = 0</th>
<th>β = 0.50</th>
<th>β = 0</th>
<th>β = 0.50</th>
<th>β = 0</th>
<th>β = 0.50</th>
<th>β = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>C = 0.35</td>
<td>-0.03</td>
<td>0.43</td>
<td>0.05</td>
<td>-0.06</td>
<td>0.42</td>
<td>0.05</td>
<td>-0.08</td>
<td>0.47</td>
<td>0.04</td>
<td>-0.11</td>
<td>0.59</td>
<td>0.04</td>
<td>-0.13</td>
<td>0.93</td>
<td>0.04</td>
</tr>
<tr>
<td>C = 0.30</td>
<td>-0.03</td>
<td>0.42</td>
<td>0.06</td>
<td>-0.06</td>
<td>0.41</td>
<td>0.05</td>
<td>-0.08</td>
<td>0.47</td>
<td>0.04</td>
<td>-0.11</td>
<td>0.59</td>
<td>0.04</td>
<td>-0.13</td>
<td>0.91</td>
<td>0.06</td>
</tr>
<tr>
<td>C = 0.25</td>
<td>-0.03</td>
<td>0.41</td>
<td>0.06</td>
<td>-0.06</td>
<td>0.41</td>
<td>0.05</td>
<td>-0.08</td>
<td>0.47</td>
<td>0.04</td>
<td>-0.11</td>
<td>0.60</td>
<td>0.04</td>
<td>-0.13</td>
<td>0.90</td>
<td>0.06</td>
</tr>
<tr>
<td>C = 0.20</td>
<td>-0.03</td>
<td>0.41</td>
<td>0.06</td>
<td>-0.06</td>
<td>0.41</td>
<td>0.05</td>
<td>-0.08</td>
<td>0.47</td>
<td>0.04</td>
<td>-0.11</td>
<td>0.60</td>
<td>0.04</td>
<td>-0.13</td>
<td>0.89</td>
<td>0.06</td>
</tr>
<tr>
<td>C = 0.15</td>
<td>-0.03</td>
<td>0.40</td>
<td>0.05</td>
<td>-0.06</td>
<td>0.41</td>
<td>0.05</td>
<td>-0.09</td>
<td>0.47</td>
<td>0.04</td>
<td>-0.11</td>
<td>0.60</td>
<td>0.04</td>
<td>-0.13</td>
<td>0.87</td>
<td>0.06</td>
</tr>
<tr>
<td>C = 0.10</td>
<td>-0.03</td>
<td>0.38</td>
<td>0.05</td>
<td>-0.06</td>
<td>0.40</td>
<td>0.05</td>
<td>-0.09</td>
<td>0.47</td>
<td>0.04</td>
<td>-0.11</td>
<td>0.60</td>
<td>0.04</td>
<td>-0.13</td>
<td>0.86</td>
<td>0.06</td>
</tr>
<tr>
<td>C = 0.05</td>
<td>-0.03</td>
<td>0.36</td>
<td>0.05</td>
<td>-0.06</td>
<td>0.39</td>
<td>0.05</td>
<td>-0.09</td>
<td>0.46</td>
<td>0.04</td>
<td>-0.11</td>
<td>0.60</td>
<td>0.04</td>
<td>-0.13</td>
<td>0.85</td>
<td>0.05</td>
</tr>
</tbody>
</table>

As indicated in

Table 5, Models C and G yielded biased ATE estimates. The bias ranged from about -0.03 to -0.13 standard deviations and increased as the pre-post correlation increased due to a complicated set of relationships in the data. The mixed approach was forced to choose between two sets of imperfect results. Overall, it closely replicated the results of the covariance adjustment approach when the cutoff was set to 0.20, which is perhaps as much as can be expected given the circumstances.

Figure 8 illustrates how the bias was introduced when the ATE was 0.50 standard deviations and the pre-post correlation was 0 and .99. When the pre-post correlation was 0, the differential treatment effect was functionally related to the pretest ability but randomly associated with the posttest ability. Thus, adding the differential treatment...
effect was like adding a normally distributed error term to the posttest distribution, affecting only its variance. The increase in variance was sufficient to induce both a floor and ceiling effect. Because the true ATE was 0.50 standard deviations, the ceiling effect for the treatment group was larger, shifting its posttest mean to the left and biasing downward the ATE estimate in the original units. The floor and ceiling effects also decreased the pooled standard deviation and reliability of the posttest, so when the underestimate in the original units was standardized and disattenuated, it was inflated somewhat. In this case, the offsetting forces of underestimation and inflation yielded an ATE estimate that on average was only about 0.03 standard deviations too small.

Figure 8. The posttest distributions for the treatment group in Scenario A4 when the pre-post correlation ($\rho$) is 0 and 0.99. Means with (*) and without (O) ceiling and floor effects are indicated, and the vertical dotted line marks the mean of the control group (ATE=0).

When the pre-post correlation was .99, the differential treatment effect was strongly related to the posttest ability such that respondents with higher posttest scores tended to benefit more from the program and those with lower posttest scores tended to be harmed more. Thus, the differential impact increased the variance of the posttest ability more that it did when the pre-post correlation was 0. This exacerbated the differential floor and ceiling effects, shifting the posttest mean of the treatment group.
Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches

Further to the left thereby increasing the downward bias. At the same time, the pooled standard deviation was greater than when the correlation was zero, so the standardized estimate was biased downward even further (the small change in reliability had little impact on the estimate).

**Results for Scenario B1: Experiments with Nonequivalence at Baseline where \( \sigma_x=\sigma_y \)**

Scenarios B1 through B4 are variations of Case B. Each incorporated a difference in treatment and control means at baseline (-0.50 standard deviations) that persisted until the time of the posttest. As discussed previously, the results of prior research suggest that the gain score approach should provide unbiased ATE estimates for Case B and the covariance adjustment approach should not. This was the case for Scenario B1 as presented in Table 6.

**Table 6**

*Simulation Results for Scenario B1: Bias, Power, and Type I Error Rates for the Covariance Adjustment, Gain Score, and Mixed Approaches*

<table>
<thead>
<tr>
<th>( p(\theta^<em>_1, \theta^</em>_2) )</th>
<th>( \beta=0.50 )</th>
<th>( \beta=0.25 )</th>
<th>( \beta=0.50 )</th>
<th>( \beta=0.25 )</th>
<th>( \beta=0.50 )</th>
<th>( \beta=0.25 )</th>
<th>( \beta=0.50 )</th>
<th>( \beta=0.25 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariance Adjustment</td>
<td>Bias</td>
<td>Power</td>
<td>Bias</td>
<td>Power</td>
<td>Bias</td>
<td>Power</td>
<td>Bias</td>
<td>Power</td>
</tr>
<tr>
<td>Model C</td>
<td>-0.50</td>
<td>0.05</td>
<td>0.42</td>
<td>-0.38</td>
<td>0.07</td>
<td>0.28</td>
<td>-0.28</td>
<td>0.14</td>
</tr>
<tr>
<td>Mixed</td>
<td>( c = 0.35 )</td>
<td>-0.05</td>
<td>0.26</td>
<td>0.15</td>
<td>0.04</td>
<td>0.31</td>
<td>0.11</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>( c = 0.30 )</td>
<td>-0.03</td>
<td>0.26</td>
<td>0.13</td>
<td>-0.02</td>
<td>0.31</td>
<td>0.10</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>( c = 0.25 )</td>
<td>-0.01</td>
<td>0.26</td>
<td>0.11</td>
<td>-0.01</td>
<td>0.32</td>
<td>0.08</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>( c = 0.20 )</td>
<td>0.00</td>
<td>0.26</td>
<td>0.09</td>
<td>0.00</td>
<td>0.32</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>( c = 0.15 )</td>
<td>0.00</td>
<td>0.26</td>
<td>0.08</td>
<td>0.00</td>
<td>0.31</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>( c = 0.10 )</td>
<td>0.01</td>
<td>0.25</td>
<td>0.07</td>
<td>0.00</td>
<td>0.31</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>( c = 0.05 )</td>
<td>0.01</td>
<td>0.25</td>
<td>0.06</td>
<td>0.00</td>
<td>0.31</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>Gain Score</td>
<td>Model G</td>
<td>0.01</td>
<td>0.25</td>
<td>0.05</td>
<td>0.01</td>
<td>0.31</td>
<td>0.05</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The covariance adjustment approach produced ATE estimates that tended to be biased, and the degree of bias depended on the underlying pre-post correlation. When the correlation was zero, bias was -0.50, and when the correlation was 0.99, the bias was virtually zero. Consequently, the covariance adjustment approach yielded Type I error rates that tended to be higher than the nominal rate, with higher rates associated with lower pre-post correlations. In contrast, the gain score approach yielded unbiased estimates and nominal error rates regardless of the underlying pre-post correlation, and it
was consistently more powerful than the covariance adjustment approach. For Scenario B1, the gain score approach was clearly superior.

Figure 9 illustrates why the covariance adjustment approach provided less biased ATE estimates as the pre-post correlation increased. The difference at baseline, \(E(X_1) - E(X_0)\), was -0.50 standard deviations. Although this difference was also found on the posttest, it was offset when by true ATE of 0.50 standard deviations, thereby causing \(E(Y_1)\) to equal \(E(Y_0)\). When the underlying pre-post correlation, \(\rho(\theta_{i1}, \theta_{i2}^*)\), was zero, the regression lines for the treatment and control groups estimated by Model C tended to be zero and the associated intercepts (\(\alpha_{C0}\) and \(\alpha_{C1}\)) tend to be equal. Model C estimates the ATE as \(\alpha_{C1} - \alpha_{C0}\), so over the long run estimates produced by Model C were 0.50 standard deviations too small. On the other hand Model G estimates the ATE as 
\[
\left[ E(Y_i) - E(Y_o) \right] - \left[ E(X_i) - E(X_o) \right],
\]
and over the long run this took the value of the true ATE (after adjusting for the unreliability of the measures).

\[
\rho(\theta_{i1}, \theta_{i2}^*) = 0 \quad \rho(\theta_{i1}, \theta_{i2}^*) = 0.99
\]

![Figure 9](image)

Figure 9. As the pre-post correlation increased in Scenario B1, the ATE estimate produced by the covariance adjustment approach, \(\alpha_{C1} - \alpha_{C0}\), became less biased; by contrast, the ATE estimate produced by the gain score approach, 
\[
\left[ E(Y_i) - E(Y_o) \right] - \left[ E(X_i) - E(X_o) \right],
\]
was unbiased regardless of the correlation.

When the underlying pre-post correlation was not zero, the slope of the regression lines estimated by Model C are determined by a more complicated set of factors. The slope of the regression lines can be expressed as

\[
\hat{\gamma}_C = \left( \frac{r_{sy} - r_{sx}^2 r_{st}}{1 - r_{st}^2} \right) \frac{s_y}{s_x}
\]

(13)
Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches

(P. Cohen, Cohen, West, & Aiken, 2002, p. 68). For Scenario B1, the only value that varies in Equation (13) is the observed pre-post correlation, $r_{xy}$ (the posttest-treatment correlation, $r_{xt}$, is zero because $E(Y_1) = E(Y_0)$, and $r_{xt}^2$ is relatively small and has the same expected value regardless of the pre-post correlation). Thus, when $\rho(\theta_{i1}^*, \theta_{i2}^*)$ is 0.99, the ratio contained in the left-hand parenthesis is close to 1, and $\hat{\gamma}_C$ approaches the ratio $\frac{s_y}{s_x}$, which in this case is also 1. As noted earlier, when $\hat{\gamma}_C = 1$, Models C and G produce the same estimates. This is why Model C provides less biased estimates as $\rho(\theta_{i1}^*, \theta_{i2}^*)$ increases. Note, however, how this result, which is related to bias rather than precision, depends on the ratio of the standard deviations of the pretest and the posttest.

For Scenario B1, the mixed approach produced results that were very similar to those produced by the gain score approach. One shortcoming of the mixed approach, however, is that it resulted in Type I error rates that were slightly higher than the nominal rate when the pre-post correlation was weak. Using the 0.20 cutoff, the error rate was 0.09 when the correlation was zero, but it achieved the nominal rate when the correlation was 0.75 or larger.

**Results for Scenario B2: Experiments with Nonequivalence at Baseline where $\sigma_x<\sigma_y$**

<table>
<thead>
<tr>
<th>$\beta=0$</th>
<th>$\beta=0.5$</th>
<th>$\rho(\theta_{i1}^<em>, \theta_{i2}^</em>)$</th>
<th>0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.99</th>
<th>0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.99</th>
<th>0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Covariance Adjustment</strong></td>
<td>Bias</td>
<td>Power</td>
<td>Rate</td>
<td>Bias</td>
<td>Power</td>
<td>Rate</td>
<td>Bias</td>
<td>Power</td>
<td>Rate</td>
<td>Bias</td>
<td>Power</td>
<td>Rate</td>
<td>Bias</td>
<td>Power</td>
<td>Rate</td>
<td>Bias</td>
<td>Power</td>
</tr>
<tr>
<td>Model C</td>
<td>-0.50</td>
<td>0.05</td>
<td>0.40</td>
<td>-0.35</td>
<td>0.08</td>
<td>0.25</td>
<td>-0.18</td>
<td>0.24</td>
<td>0.11</td>
<td>-0.02</td>
<td>0.64</td>
<td>0.05</td>
<td>0.15</td>
<td>1.00</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = 0.35$</td>
<td>0.00</td>
<td>0.35</td>
<td>0.11</td>
<td>0.01</td>
<td>0.40</td>
<td>0.09</td>
<td>0.02</td>
<td>0.53</td>
<td>0.07</td>
<td>0.03</td>
<td>0.77</td>
<td>0.06</td>
<td>0.05</td>
<td>1.00</td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = 0.30$</td>
<td>0.02</td>
<td>0.36</td>
<td>0.09</td>
<td>0.01</td>
<td>0.40</td>
<td>0.07</td>
<td>0.02</td>
<td>0.54</td>
<td>0.06</td>
<td>0.03</td>
<td>0.78</td>
<td>0.06</td>
<td>0.04</td>
<td>1.00</td>
<td>0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = 0.25$</td>
<td>0.03</td>
<td>0.36</td>
<td>0.07</td>
<td>0.01</td>
<td>0.41</td>
<td>0.06</td>
<td>0.03</td>
<td>0.54</td>
<td>0.06</td>
<td>0.03</td>
<td>0.78</td>
<td>0.06</td>
<td>0.04</td>
<td>1.00</td>
<td>0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = 0.20$</td>
<td>0.04</td>
<td>0.36</td>
<td>0.06</td>
<td>0.02</td>
<td>0.41</td>
<td>0.06</td>
<td>0.03</td>
<td>0.54</td>
<td>0.06</td>
<td>0.04</td>
<td>0.78</td>
<td>0.06</td>
<td>0.04</td>
<td>1.00</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = 0.15$</td>
<td>0.04</td>
<td>0.36</td>
<td>0.06</td>
<td>0.02</td>
<td>0.41</td>
<td>0.05</td>
<td>0.03</td>
<td>0.55</td>
<td>0.06</td>
<td>0.04</td>
<td>0.78</td>
<td>0.06</td>
<td>0.04</td>
<td>1.00</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = 0.10$</td>
<td>0.04</td>
<td>0.36</td>
<td>0.06</td>
<td>0.02</td>
<td>0.41</td>
<td>0.05</td>
<td>0.03</td>
<td>0.55</td>
<td>0.06</td>
<td>0.04</td>
<td>0.78</td>
<td>0.06</td>
<td>0.04</td>
<td>1.00</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = 0.05$</td>
<td>0.04</td>
<td>0.36</td>
<td>0.05</td>
<td>0.02</td>
<td>0.41</td>
<td>0.05</td>
<td>0.03</td>
<td>0.55</td>
<td>0.06</td>
<td>0.04</td>
<td>0.78</td>
<td>0.06</td>
<td>0.04</td>
<td>1.00</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gain Score</strong></td>
<td>Bias</td>
<td>Power</td>
<td>Rate</td>
<td>Bias</td>
<td>Power</td>
<td>Rate</td>
<td>Bias</td>
<td>Power</td>
<td>Rate</td>
<td>Bias</td>
<td>Power</td>
<td>Rate</td>
<td>Bias</td>
<td>Power</td>
<td>Rate</td>
<td>Bias</td>
<td>Power</td>
</tr>
<tr>
<td>Model G</td>
<td>0.04</td>
<td>0.36</td>
<td>0.05</td>
<td>0.02</td>
<td>0.41</td>
<td>0.05</td>
<td>0.03</td>
<td>0.55</td>
<td>0.06</td>
<td>0.04</td>
<td>0.78</td>
<td>0.06</td>
<td>0.04</td>
<td>1.00</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The structure of the data for Scenario B2 was the same as that for Scenario B1 except that the standard deviation of the true post-intervention ability distribution was increased from 1 to \( \sqrt{2} \). This small difference negatively affected the bias of ATE estimates produced by both approaches. For the covariance adjustment approach, bias was still associated with the strength of the pre-post correlation, but the range of that bias was greater (-0.50 standard deviations when the pre-post correlation was 0, up to 0.15 standard deviations when the correlation was 0.99). Just as in Scenario B1, the expected value of the slope, \( \hat{\gamma}_c \), approached the ratio of the standard errors as \( r_{xy} \) approached 1. In this Scenario, however, the ratio was \( \sqrt{2} \). Thus, the covariance adjustment approach produced (essentially) unbiased ATE estimates when \( \rho(\theta_{i1}, \theta_{i2}) \) was 0.75 (roughly \( \frac{1}{\sqrt{2}} \)), and stronger pre-post correlations were associated with positively biased estimates.

![Figure 10](image-url)

Figure 10. A distribution of posttest sum scores for Scenario B2 exhibiting simultaneous floor and ceiling effects (solid line) compared to a distribution of posttest sum scores for Scenario B1 without floor or ceiling effects (dashed line).

For both approaches, increasing the standard deviation of the post-intervention ability distribution led to another source of bias. The range of item difficulties on the posttest was insufficient to encompass the full posttest ability distribution. This caused the range of the observed sum scores to be restricted because of simultaneous floor and ceiling effects (Figure 10). While the resultant decrease in the pooled standard deviation
was small, it over-inflated ATE estimates enough to be noticed. For estimates produced by the gain score approach, this meant that standardizing the impact estimates transformed them from being unbiased in the original units to being slightly biased when standardized (about 0.04 standard deviations).

Overall, the gain score approach is the lesser of two evils. The bias was small, power was greater, and the Type I error rate was at or near nominal levels. The mixed approach, by and large, reflected this preference, yielding results that were almost identical to those of the gain score approach.

Results for Scenario B3: Experiments with Nonequivalence at Baseline and Extreme Growth

As with Scenario A3, the results for Scenario B3 were not ideal because a ceiling effect was induced by substantial growth over time. The degree of bias in the ATE estimates produced by the covariance adjustment approach was a function of the pre-post correlation, while the bias in estimates produced by the gain score approach was not. However, in the latter case the bias was consistent and material (about 0.35 standard deviations). The Type I error rate for both approaches was a function of the correlation. The pattern for the covariance adjustment approach was similar to the one observed for Scenario B2—the error rate decreased as the correlation increased, and then it increased again when correlation was .99. For the gain score approach, the error rate increased with the correlation. The mixed approach with a cutoff of 0.20 yielded results that closely replicated those of the gain score approach. Overall, the results for all three approaches were problematic.

Table 8
Simulation Results for Scenario B3: Bias, Power, and Type I Error Rates for the Covariance Adjustment, Gain Score, and Mixed Approaches

<table>
<thead>
<tr>
<th>β = 0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model C</td>
<td>-0.51</td>
<td>0.05</td>
<td>0.37</td>
<td>-0.35</td>
</tr>
<tr>
<td>Mixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c = 0.35</td>
<td>0.32</td>
<td>0.49</td>
<td>0.12</td>
<td>0.34</td>
</tr>
<tr>
<td>c = 0.30</td>
<td>0.33</td>
<td>0.49</td>
<td>0.11</td>
<td>0.34</td>
</tr>
<tr>
<td>c = 0.25</td>
<td>0.33</td>
<td>0.49</td>
<td>0.10</td>
<td>0.34</td>
</tr>
<tr>
<td>c = 0.20</td>
<td>0.33</td>
<td>0.49</td>
<td>0.10</td>
<td>0.35</td>
</tr>
<tr>
<td>c = 0.15</td>
<td>0.34</td>
<td>0.49</td>
<td>0.10</td>
<td>0.35</td>
</tr>
<tr>
<td>c = 0.10</td>
<td>0.34</td>
<td>0.49</td>
<td>0.09</td>
<td>0.35</td>
</tr>
<tr>
<td>c = 0.05</td>
<td>0.34</td>
<td>0.49</td>
<td>0.09</td>
<td>0.35</td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model G</td>
<td>0.34</td>
<td>0.49</td>
<td>0.09</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches
Figure 11 illustrates how the ceiling effect leads to these results. The difference between the pretest means of the treatment and control groups is $\lambda$, which is standardized with respect to the posttest distribution. In this case, the treatment pretest distribution was shifted $\lambda$ standard deviations to the left of the control pretest distribution. The pretest and posttest distributions were separated by $\phi$, the average growth (in the absence of the treatment) between the time of the pretest and posttest.

The ceiling effect caused the control group’s posttest mean to be shifted $a$ units to the left, and when the true ATE was zero it caused the treatment group’s posttest mean to be shifted $d$ units to the left. Because $d < a$, the models yielded an erroneous positive ATE estimate in the original units of the measure when the null hypothesis was true. This
Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches

compromised the Type I error rate. The estimate of the standardized effect was inflated further by the reduced pooled standard deviation and reliability that result from the posttest ceiling effect. For the gain score approach, Model G became more sensitive to the erroneous non-zero effect as the pre-post correlation increased. When the treatment produced an ATE of 0.5 standard deviations, the pretest and posttest means were shifted to the left by the same amount, \( a \) units. This led to a correct estimate of the ATE in the original units of the measure, but an inflated estimate when standardized. This illustrates a circumstance in which standardization has the potential to distort ATE estimates and it suggests that researchers should report impact estimates in original and standardized units.

Results for Scenario B4: Experiments with Nonequivalence at Baseline with Differential Impacts

When the assumption of an average treatment effect is unfounded, neither Model G nor C is the correct model. In this scenario, the covariance adjustment approach provided biased estimates, and again the bias and Type I error rates were a function of the pre-post correlation. The gain score approach provided a constant but biased estimate over the entire range of correlations, but with error rates that are at roughly the nominal level. While power for both approaches increased as the pre-post correlation increased, the gain score approach was not universally more powerful as it was in Scenarios B1-B3. As in Scenarios B1-B3, the mixed approach replicates the results of the gain score approach.

Table 9
Simulation Results for Scenario B4: Bias, Power, and Type I Error Rates for the Covariance Adjustment, Gain Score, and Mixed Approaches

<table>
<thead>
<tr>
<th>( \rho^{0}<em>{\theta</em>{i1}, \theta_{i2}} )</th>
<th>0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta=0.50 )</td>
<td>Type I Error</td>
<td>Bias</td>
<td>Power</td>
<td>Rate</td>
<td>Bias</td>
</tr>
<tr>
<td>( \beta=0 )</td>
<td>Model C</td>
<td>-0.52</td>
<td>0.05</td>
<td>0.41</td>
<td>-0.37</td>
</tr>
<tr>
<td>Mixed</td>
<td>c = 0.35</td>
<td>-0.18</td>
<td>0.20</td>
<td>0.11</td>
<td>-0.17</td>
</tr>
<tr>
<td>Mixed</td>
<td>c = 0.30</td>
<td>-0.17</td>
<td>0.20</td>
<td>0.09</td>
<td>-0.16</td>
</tr>
<tr>
<td>Mixed</td>
<td>c = 0.25</td>
<td>-0.16</td>
<td>0.21</td>
<td>0.08</td>
<td>-0.15</td>
</tr>
<tr>
<td>Mixed</td>
<td>c = 0.20</td>
<td>-0.16</td>
<td>0.21</td>
<td>0.07</td>
<td>-0.15</td>
</tr>
<tr>
<td>Mixed</td>
<td>c = 0.15</td>
<td>-0.15</td>
<td>0.21</td>
<td>0.06</td>
<td>-0.15</td>
</tr>
<tr>
<td>Mixed</td>
<td>c = 0.10</td>
<td>-0.15</td>
<td>0.20</td>
<td>0.06</td>
<td>-0.15</td>
</tr>
<tr>
<td>Mixed</td>
<td>c = 0.05</td>
<td>-0.15</td>
<td>0.21</td>
<td>0.06</td>
<td>-0.15</td>
</tr>
<tr>
<td>Gain Score</td>
<td>Model G</td>
<td>-0.15</td>
<td>0.21</td>
<td>0.06</td>
<td>-0.15</td>
</tr>
</tbody>
</table>

32
Researchers who use the two-group pretest-posttest design must choose to analyze their data with the covariance adjustment approach or the gain score approach. Prior research suggests that when the underlying structure of the data conforms to what I labeled Case A, the former should be preferred because it affords greater statistical power. When the underlying structure conforms to Case B, the latter should be preferred because it reduces bias. It is unlikely that researchers know with certainty the case with which they are confronted, especially if they implemented a randomized trial in a field setting and suspect that their design may be compromised to some extent. A number of researchers have suggested that in circumstances such as these, data analysts should inspect their data and use their judgment to choose an appropriate approach. However, the structure of the observed data can be quite complex, especially if one takes into account (a) the relationship between unobservable latent variables of interest and the imperfect observable measures researchers use to estimate them, and (b) the effect of converting ATE estimates to disattenuated standardized effects. Visual inspection and judgment, while an important part of data analysis, is not enough.

I have suggested a simple, objective rule for choosing between approaches. This mixed approach compares $\hat{\lambda}$, an estimate of the bias that would occur if Case B holds, to $c$, a predetermined cutoff value. When $\hat{\lambda} > c$, the gain score approach is used, otherwise the covariance adjustment approach is used. Based on the results of the simulations presented here, I have suggested setting $c$ to 0.20 standard deviations. Although it remains to be seen if this is a useful suggestion for data structures that were not investigated in this study, the exact choice of $c$ may not be crucial because in many circumstances investigated here results were improved for values of $c$ that ranged from 0.10 to 0.30 standard deviations.

One way to judge the mixed approach is by its results in each of the eight data scenarios. The mixed approach was successful when the data conformed to paradigmatic conditions (Scenarios A1, A2, and B1) or when ceiling and floor effects were small (Scenarios A3 and B2). In these five scenarios, either the covariance adjustment or gain score approach was clearly preferred, and a moderate pre-post correlation of 0.75 (and sometimes less) was enough to ensure that the results obtained using the mixed approach were in line with those of the preferred approach. The mixed approach was less successful when there were large ceiling and floor effects (Scenario B3), the estimate of an average treatment effect was not appropriate (Scenario A4), or both (Scenario B4). In these three scenarios, neither the covariance adjustment nor gain score approach consistently produced unbiased results. The mixed approach fell short because it was forced to choose between two imperfect options, although it made its choice as past research suggests it should by choosing the covariance adjustment approach when Case A held and the gain score approach when Case B held.

Another way to judge the mixed approach is by looking at particular circumstances in which researchers must determine whether Case A or B holds. If the possibilities were limited to Scenario A1 versus B1, the penalty for using the mixed approach would be a trivially small reduction in power and a small increase in the Type I
error rate (both of which decrease as the pre-post correlation increases). If the possibilities were limited to Scenario A2 versus B2, the penalty would be even smaller. If the possibilities were limited to A3 versus B3, or A4 versus B4, the mixed approach would not perform as well. In the former situation, inflated Type I error rates and high bias would likely undermine results, regardless of approach. In the latter situation, the mixed approach may offer a less biased result, albeit at the cost of diminished power and increased Type I error.

Table 10
*The Estimated Penalty Researchers Pay When They Must Determine if Case A or B Holds*

<table>
<thead>
<tr>
<th>Observed Range (Across Cases A and B)</th>
<th>Possible Scenarios</th>
<th>Bias</th>
<th>Loss in Power</th>
<th>Type I Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A1 or B1</td>
<td>0.00</td>
<td>almost none</td>
<td>.04 to .09</td>
</tr>
<tr>
<td></td>
<td>A2 or B2</td>
<td>0.00 to 0.06</td>
<td>none</td>
<td>.05 to .06</td>
</tr>
<tr>
<td></td>
<td>A3 or B3</td>
<td>-0.03 to 0.35</td>
<td>substantial</td>
<td>.05 to .37</td>
</tr>
<tr>
<td></td>
<td>A4 or B4</td>
<td>-0.03 to -0.16</td>
<td>substantial</td>
<td>.04 to .07</td>
</tr>
</tbody>
</table>

The implication of these results is that the mixed approach is more appropriate when (a) ceiling and floor effects are small, and (b) the estimates of an ATE as operationalized in Models C and G is appropriate. The good news is that researchers can know the distribution of pretest and posttest scores, so they can determine whether these warrant concern. And they can fit a set of plausible models (Burnham & Anderson, 2002) to help establish that it is reasonable to estimate an ATE and ignore potential interactions. What data analysts cannot observe is the true data generating mechanism. So given that score distributions and statistical models are not problematic, the results of this investigation suggest that the mixed approach can help researchers judge more usefully what they cannot directly observe.

These results also confirm that the relative reliability of outcome measures (gain scores versus residualized gain scores) does not provide a useful way to choose an approach. For example, Zumbo’s (1999) flow chart (Figure 1), which is premised on maximizing the reliability of the outcome measure, offers a point of comparison. The flow chart indicates that we should always use the covariance adjustment approach for Scenarios A1 and B1. Although this is acceptable for Scenario A1, for Scenario B1 it would neither maximize power nor minimize bias. And for Scenarios A2 and B2, Zumbo’s flow chart suggests that the choice of approach should depend on the strength of the pre-post correlation, which is not supported by the simulation results.
Finally, the complications introduced by simultaneous floor and ceiling effects suggest that we should be cautious when considering standardized effect sizes. One advantage of standardizing an ATE estimate is that it can be easily compared to standardized estimates from other studies. However, it has the potential to bias an estimate if the distribution of the outcome measure is truncated. It would seem prudent, therefore, to report impact estimates in the original units in addition to standardized units, especially when experimental designs may have been compromised.

While the mixed approach may offer some improvement in a restricted set of circumstances, there are countless ways that an experiment can go wrong. Consequently, the results reported here should be considered as being only preliminary. It remains to be seen if combining the mixed approach with other corrective measures such as propensity score matching would also lead to improved ATE estimates. And these results may not hold for combinations of sample size and effect size that were not investigated here. At the very least, it offers researchers a way to choose an approach using a systematic rule that works in some circumstances rather than relying solely on the unevaluated intuition and judgment of individual data analysts.
Appendix A: S-Plus Code Used in the Simulations

The simulations were executed using the function \texttt{gainsim} written by the author. The function requires the following input values:

- \texttt{mypath}: the path for the output file as a string in the form "drive:\folder\subfolder\...\"
- \texttt{myseed}: the seed for the random number generator
- \texttt{ntrial}: the number of iterations for each pre-post correlation (default is 3000)
- \texttt{ncase}: the total sample size (treatment + control) for each study (default is 60)
- \texttt{nitem}: the number of items on the pretest and posttest (default is 60)
- \texttt{max.delta}: the maximum value in logits for the delta item difficulty parameters (default is 2)
- \texttt{corpp}: a vector containing the unadjusted correlations of the pre and post ability distributions (default vector is 0, .25, .5, .75, .99).
- \texttt{tau}: standard deviation of the posttest ability distribution
- \texttt{phi}: growth in abilities that is common to treatment and control groups
- \texttt{lambda}: difference at baseline used to create Case B
- \texttt{beta}: true ATE
- \texttt{omega}: a treatment by pretest ability interaction

As \texttt{gainsim} runs, it calls seven subfunctions. All were written by the author with the exception of \texttt{cralpha}. The subfunctions are:

- \texttt{cralpha}: calculates Cronbach's alpha (author unknown)
- \texttt{delta.vect}: creates a vector of item difficulties
- \texttt{mat.maker}: creates a matrix to store output values
- \texttt{ppnorm}: draws a random sample of true pre and post ability estimates
- \texttt{randass}: randomly assigns subjects to treatment and control conditions
- \texttt{rashcdata}: simulates item responses for a test that conforms to the Rasch model
- \texttt{spooled}: calculates the pooled standard deviation

The code for each is given below.
gainsim<-function(mypath, myseed, ntrial=3000, ncase=60, nitem=60, max.delta=2, corpp=c(0, .25, .5, .75, .99), tau, phi, lambda, beta, omega)
{
  # set the random seed
  set.seed(myseed)

  # Transform the standardized effect sizes to the units
  # of true ability distribution at posttest
  beta<-beta*tau
  omega<-omega*tau
  lambda<-lambda*tau
  phi<-phi*tau

  # Set mean and sd of the unadjusted abilities
  mmat<-c(0,phi)
  sdmat<-c(1,tau)

  # Compute the number of correlation values to be
  # considered for each data scenario.
  ncor<-length(corpp)

  # Create the vector of item difficulties.
  delta<-delta.vect(nitem, max.delta)

  # Iterate over each unadjusted pre-post correlation
  for (j in 1:ncor)
  {
    # Track the iterations by printing the count
    # to the screen
    print(c(j, date()))

    # Create a temporary output matrix
"stesgdis", "stesppdis", "coefg", "seg", "tg", "pvalg", "coefpp", "sepp", "tpp", "pvalpp", "diffpre")

# Iterate over the number of datasets to be simulated.
for (i in 1:ntrial)
{
    # Randomly draw pre and post true abilities
    pp<-ppnorm(ncase, mmat[1], sdmat[1], mmat[2],
                sdmat[2], corpp[j])

    # Randomly assign units to treatment and
    # control conditions.
    trt<-randass(ncase/2, ncase/2)

    # adjust pretest scores
    pp[,1]<-pp[,1]+lambda*trt

    # adjust posttest scores
    pp[,2]<-pp[,2] + trt*( omega*(pp[,1]-mean(pp[,1])
                            + lambda + beta )

    # Create pre and post response matrices based
    # on the Rasch model.
    resp1<-raschdata(pp[,1], delta)
    resp2<-raschdata(pp[,2], delta)

    # Create pretest, posttest, and gain scores as
    # sum scores
    pp[,1]<- rowSums(resp1)
    pp[,2]<- rowSums(resp2)
    gain<-pp[,2]-pp[,1]

    # Fit model C
    lmpp<-summary(lm(pp[,2]~trt+pp[,1]))
    lmpp<-lmpp$coefficients["trt", ]

    # Fit model G
    lmg<-summary(lm(gain~trt))
    lmg<-lmg$coefficients["trt", ]

    #Calculate the reliability of the test scores
    relpre<-cralpha(resp1)
    relpost<-cralpha(resp2)
    relgain<-cralpha(resp2-resp1)
# Calculate the observed pooled standard deviation
spoolp<-spooled(pp[,2], trt)

# Calculate the disattentuated pooled standard deviation
spoolpdis<-spoolp*sqrt(relpost)

# Enter one row of the temporary output matrix
outmat[i,c("seed", "ntrial", "rho", "lambda", "beta", "tau", "phi", "omega")]<-c(myseed, ntrial, corpp[j], lambda, beta, tau, phi, omega)
outmat[i,"vgain"]<-var(gain)
outmat[i,"vpre"]<-var(pp[,1])
outmat[i,"vpost"]<-var(pp[,2])
outmat[i,"spoolp"]<-spoolp
outmat[i,"spoolpdis"]<-spoolpdis
outmat[i,"relpre “]<-relpre
outmat[i,"relpost “]<-relpost
outmat[i,"relgain “]<-relgain
outmat[i,"rpp “]<-cor(pp)[1,2]
outmat[i,"stesg"]<-lmg[1]/spoolp
outmat[i,"stespp"]<-lmpp[1]/spoolp
outmat[i,"stesgdis"]<-lmg[1]/spoolpdis
outmat[i,"stesppdis"]<-lmpp[1]/spoolpdis
outmat[i,c("coefg", "seg", "tg", "pvalg")]<-t(lmg)
outmat[i,c("coefpp", "sepp", "tpp", "pvalpp")]<-t(lmpp)
outmat[i,"diffpre “]<-diff(tapply(pp[,1], trt, mean))

# Export the temporary data file
file.name<-paste(mypath, "\gain_", myseed, ".txt", sep="")
write.table(outmat, file = file.name, sep = "\t", append = T, quote.strings = F, dimnames.write = (file.exists(file.name)[1]==FALSE) )

# Print time the simulation ended.
print(date())
Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches

########################################################
##                       CRALPHA                      ##
##   Calculates Cronbach's alpha. User inputs a matrix of observed item scores. (Based on code written by another unknown author.)
########################################################

cralpha<-function (mydata)
{
    nitem<-ncol(mydata)
    if (length(dim(mydata))==2) mydata<-array(mydata,
    
    dim=c(nrow(mydata), ncol(mydata), 1))
    nsets<-dim(mydata)[3]
    relmat<-matrix(NA, nrow=nsets, ncol=1)
    for (i in 1:nsets)
    {
        relmat[i,]<-sum(colVars(mydata[,,i]))
        relmat[i,]<-relmat[i]/ var(rowSums(mydata[,,i]))
        relmat[i,]<-1-relmat[i,]
        relmat[i,]<-relmat[i] * nitem/(nitem-1)
    }
    return(relmat)
}

########################################################
##                     DELTA.VECT                     ##
##   Creates a perfectly uniform distribution of item difficulties that are symmetric about zero and range from -max.delta to max.delta. User inputs the number of items (nitem) and the maximum delta value in logits (maxdelta).
########################################################
delta.vect<-function(nitem, max.delta)
{
    return ( ( c(0:(nitem-1))-(((nitem-1)/2) ) / ((nitem-1)/(2*max.delta)) ) )
}
Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches

mat.maker<-function(nrow, col.names)
{
ncol<-length(col.names)
mymat<-matrix(NA, ncol=ncol, nrow=nrow)
dimnames(mymat)<-list(NULL, col.names)
return(mymat)
}

ppnorm<-function(n, mnpre, sdpre, mnpost, sdpost, rho)
{pp<-(rmvnorm(n, mean=c(mnpre, mnpost), sd=c(sdpre, sdpost), rho=rho))
dimnames(pp)<-list(NULL, c("pre", "post"))
return(pp)
}

randass<-function(ntrt, nctl)
{
n <- ntrt + nctl
return(matrix(c(rep(1, ntrt), c(rep(0, nctl))))[order(runif(n, 0, 1)), ])
}
Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches

########################################################
##                       RASCHDATA                    ##
##   Randomly generates dichotomous item responses    ##
##   that conform to the Rasch model. User inputs    ##
##   a vector of abilities (theta) and a vector of    ##
##   item difficulties (delta)                        ##
########################################################

raschdata<-function(theta, delta)
{
  ncase <- length(theta)
  nitem <- length(delta)
  respmat <- matrix(NA, nrow = ncase, ncol = nitem)
  delta <- t(matrix(rep(t(delta), ncase), ncol = ncase))
  respmat <- exp(theta - delta)
  respmat <- respmat/(1 + respmat)
  respmat <- matrix(rbinom(ncase * nitem, 1, respmat), nrow = ncase)
  return(respmat)
}

########################################################
##                       SPOOLED                      ##
##   Calculates the pooled standard deviation. User   ##
##   inputs the a vector (y) and an indicator (0/1)   ##
##   of treatment/control condition (trt).            ##
########################################################

spooled<-function(y, trt)
{
  spool<-tapply(y, trt, var)
  spool<-c(sum(1-trt)-1, sum(trt)-1)*spool
  spool<-sum(spool)
  spool<-spool/(length(trt)-2)
  spool<-sqrt(spool)
  return(sqrt(spool))
}
Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches

References


Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches


Paper 1: Choosing Between Gain Score and Covariance Adjustment Approaches


A Reparameterized Rasch Model

Introduction

This paper describes a reparameterized Rasch model (RRM) that can be fit as a hierarchical generalized linear model (HGLM) in order to obtain estimates of item group difficulties and associated tests of significance. At the same time, the RRM produces the item and person estimates provided by a traditionally parameterized Rasch model. This information can be useful to evaluators in many ways, and in this paper I explore how it can help them validate a measurement scale and test the homogeneity of respondents in relationship to groups of items.

In order to use the RRM, two conditions must be met. First, the items that comprise a measure must fall into mutually exclusive groups that differentiate the items on features of interest. For example, a researcher might group items according to the sub-parts of the construct being measured (e.g., the logical structure of the items or the educational standards being addressed); the construct or target (e.g., life at home versus life at work) that one is asked to consider when responding to survey questions; or the degree of presumed alignment between items and the program. These groupings are not intended to represent distinct dimensions, such as math and language abilities; rather, they are intended to divide a single dimension in a useful way. Second, there must be a practical interpretation of differences or similarities across item groups. The difficulty of items can vary for many reasons. In order for a comparison of item groups to be interpretable, a researcher needs a rationale for why subjects respond to items as they do, and why this will be reflected in differences or similarities across item groups.

In the sections that follow, I present three models—the traditionally parameterized Rasch model, the reparameterized Rasch model, and the multidimensional random coefficients multinomial logit model. I discuss the distributional assumptions of the models and describe how each can be fit as a hierarchical generalized linear model. Then, I provide an example that illustrates how the RRM can be applied by researchers and evaluators.

Statistical Models

The Rasch Model

In the literature associated with item response theory, the simple Rasch model is often presented as

$$\eta_{ij} = \theta_j - \delta_i,$$

(1)

where $\theta_j$ is the trait level of person $j$ and $\delta_i$ is the difficulty of item $i$ (De Boeck & Wilson, 2004; Embretson & Reise, 2000; Rasch, 1960/1980). The dependent variable, the logit $\eta_{ij}$, is the log of the odds that person $j$ will realize a score of 1 (versus 0) on item $i$. A score of 1 can represent any number of things, such as answering a question...
correctly on a test, agreeing with a statement on a survey, or successfully completing a task. The logit can be rewritten more explicitly as

$$\eta_{ij} = \ln \left( \frac{P(Y_{ij} = 1 | \theta_j, \delta_i)}{1 - P(Y_{ij} = 1 | \theta_j, \delta_i)} \right),$$

where $Y_{ij}$ is the score produced by person $j$ on item $i$. Equation (2) can be rearranged as the conditional probability

$$P(Y_{ij} = 1 | \theta_j, \delta_i) = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}.$$

Item responses are treated as Bernoulli distributed random variables with the probability of success given by Equation (3), which depends on the trait level and item difficulty.

The Rasch model can be fit as a two-level hierarchical generalized linear model (HGLM) with a logit link (Adams, Wilson, & Wu, 1997; Kamata, 2001; Raudenbush & Bryk, 2002, pp. 365-368) using a variety of software packages (Tuerlinckx et al., 2004). The examples in this paper were fit using HLM 6 (Raudenbush, Bryk, & Congdon, 2008). The combined HGLM for an instrument with $M$ items takes the form

$$\eta_{ij} = \theta_j + \delta_0 + \sum_{i=1}^{M-1} \delta'_i I_i.$$

Here, $I_i$ is an effect-coded variable indicating which of the $M$ items is associated with the score produced by person $j$. The model incorporates $M-1$ effect-coded variables because one variable (item $M$ in this case) is used as a reference item. In this model, the $\delta_i$ from Equation (1) has been partitioned into $\delta_0$ (the intercept) and $\delta'_i$ (a deviation). Note that the sign associated with these two values is now positive because of the signs associated with the effect-coded variables. By changing the signs, the intercept can be interpreted in different ways: $-\delta_0$ is the average item difficulty, $\delta_0$ is the average log odds of success for the $M$ items, and $\exp(\delta_0) / (1 + \exp(\delta_0))$ is the base rate of success across all items. The t-test associated with $\delta_0$ can be applied to any and all of these interpretations to determine if the average item difficulty differs from 0, the log odds of success differs from 0, or the base rate of success differs from 50%.
The $\delta_i$ from Equation (1) can be calculated from the $\delta'_i$'s and $\delta_0$ in Equation (4). Because the $\delta'_i$'s are expressed as deviations from $\delta_0$ and the sum of the $\delta'_i$'s is constrained to be zero, the item difficulties can be computed as

$$\delta_i = -\left(\delta_0 + \delta'_i\right) \quad \text{for nonreference items (i = 1 to M-1), and}$$

$$\delta_M = -\left(\delta_0 - \sum_{i=1}^{M-1} \delta'_i\right) \quad \text{for the reference item.}$$

Note how the sign has been reversed in order to obtain estimates of item difficulty. Breaking the combined model out by level, the level-one model is

$$\eta_{ij} = \pi_{0j} + \sum_{i=1}^{M-1} \pi_{ij} I_i,$$  

and the level-2 model is

$$\begin{align*}
\pi_{0j} &= \delta_0 + \theta_j \\
\pi_{ij} &= \delta'_i \\
\vdots \\
\pi_{(M-1)j} &= \delta'_{(M-1)}. 
\end{align*}$$

The trait level, $\theta_j$, is treated as a random effect. In what is perhaps a more common notation presented in Raudenbush and Bryk (2002), $\theta_j$ is denoted by $u_{0j}$, the error term associated with $\pi_{0j}$, which is assumed to be normally distributed with a mean of 0 and a constant variance ($\tau_{00}$). The $\delta'_i$'s and $\delta_0$ are treated as fixed effects that represent the intercepts associated with $\pi_{0j}$ through $\pi_{(M-1)j}$ (in Raudenbush and Bryk’s notation they are denoted $\beta_{00}$ through $\beta_{(M-1)0}$).

Table 1 provides an example of a level-1 data file for HLM 6 for a six-item instrument with five effect-coded indicator variables ($I_j$ through $I_5$). Item indicators are sometimes dummy coded (see for example Raudenbush, Johnson, & Sampson, 2003), but for our purposes they are effect coded: -1 for the reference item, 1 for a non reference item associated with the given response, and 0 otherwise. Note that item six is the reference item, the data relate to only two respondents, and the item numbers are included for clarity and are not required.
Table 1
A Hypothetical HLM 6 Level-1 Data File That Could Be Used to Fit a Traditionally Parameterized Rasch Model

<table>
<thead>
<tr>
<th>subject</th>
<th>item</th>
<th>response</th>
<th>effect coded variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>j</td>
<td>i</td>
<td>y_{ij}</td>
<td>I_1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>

The RRM

The Rasch models presented in Equations (1) and (4) include \( M-1 \) fixed effects (associated with item difficulties) and \( l \) random effect (associated with person abilities). The RRM described here has the same number of fixed and random effects, however, the \( M-1 \) effect-coded variables are partitioned in a way that allows item-group parameters to be estimated. Building on the HGLM presentation, the RRM can be expressed as

\[
\eta_j = \theta_j + \gamma_0 + \sum_{g=1}^{G-1} \gamma_g' X_g + \sum_{i=0}^{M} \delta' I_i. \tag{8}
\]

Here, \( g \) indexes the item groups, ranging from 1 to \( G \). The variable \( X_g \) is effect coded and identifies the group with which item \( i \) is associated. Breaking the combined model out by level, the level-one model is

\[
\eta_j = \pi_{0j} + \sum_{g=1}^{G-1} \pi_g X_g + \sum_{i=0}^{M} \pi_i I_i. \tag{9}
\]
The level-2 model is

\[
\begin{align*}
\pi_{0j} &= \gamma_0 + \theta_j \\
\pi_{1j} &= \gamma'_1 \\
& \quad \ldots \\
\pi_{(G-1)j} &= \gamma'_{(G-1)} \\
\pi_{Gj} &= \delta'_G \\
& \quad \ldots \\
\pi_{(M-1)j} &= \delta'_M.
\end{align*}
\]  

As before, the variables that identify the items associated with each response, \(I_i\), are effect coded, but they are coded separately within item groups. Consequently, only \(M-G\) item-level identifiers are entered into the model. In Equation (8) these correspond to items with the subscript \(i\) ranging from \((G+1)\) to \(M\). The remaining \(G\) items are reference items. In Equation (8) these correspond to items with the subscript \(i\) ranging from \(1\) to \(G\), and for convenience item \(i\) is taken to be a member of group \(g\) when \(i=g\). Table 2 (A) illustrates this coding where items 1, 3, and 4 comprise item group 1, and items 2, 5, and 6 comprise item group 2.

It is important to note that the assignment of the subscript \(i\) to items in Equation (8) is arbitrary. All that matters in practice is that items are effect coded within item groups and that one item per group is used as a reference item. Table 2 (B) illustrates an alternate and equally valid effect coding for the same six items.
Table 2

*A Hypothetical HLM 6 Level-1 Data File That Could Be Used to Fit an RRM*

(A) The assignment of subscript $i$ to items corresponds to Equation (8)

<table>
<thead>
<tr>
<th>subject</th>
<th>item</th>
<th>response</th>
<th>$y_{ij}$</th>
<th>$X_1$</th>
<th>$I_3$</th>
<th>$I_4$</th>
<th>$I_5$</th>
<th>$I_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

(B) An alternative assignment of subscript $i$ to items

<table>
<thead>
<tr>
<th>subject</th>
<th>item</th>
<th>response</th>
<th>$y_{ij}$</th>
<th>$X_1$</th>
<th>$I_1$</th>
<th>$I_3$</th>
<th>$I_2$</th>
<th>$I_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td></td>
</tr>
</tbody>
</table>

The intercept of the RRM is denoted as $\gamma_0$, and it has a different interpretation from $\delta_0$, the intercept for the traditionally parameterized Rasch model presented in Equation (4). For the RRM, $-\gamma_0$ is the average difficulty of the item groups (rather than the items), $\gamma_0$ is the average log odds of success across item groups, and
exp(γ₀)/(1 + exp(γ₀)) is the base rate of success across groups. When the data are balanced (every item group has the same number of items), γ₀ will equal δ₀; when the data are unbalanced, this may not be the case.

The difficulty of an item group is the average difficulty of the items within the group, and it can be calculated as

\[ γ_g = -(γ_0 + γ_g') \quad \text{for nonreference item groups (g = 1 to G-1), and} \]

\[ γ_G = -\left(γ_0 - \sum_{g=1}^{G-1} γ_g'\right) \quad \text{for the reference item group.} \]  

The item difficulty parameters can then be calculated as

\[ δ_i = γ_{g+i} - δ_i' \quad \text{for nonreference items in group } g \]

\[ δ_k = γ_{g+k} + \sum_{i \neq g} δ_i' \quad \text{for the reference item in group } g \]  

Because item groups are mutually exclusive, the item difficulty estimates produced by the RRM are, in theory, the same as those produced by the traditionally parameterized Rasch model. In practice, the estimates may differ to a small degree depending on the precision of the estimation procedure.

The level-2 model can be extended to include explanatory covariates. In the level-2 model given by Equations (13), the person-level covariate \( Z_j \) is included as a main effect associated with the average item difficulty, and as an interaction associated with item groups. This is an example of what is sometimes called a person explanatory model because a characteristic of a person is used to explain item responses (Wilson & De Boeck, 2004). In this case, the level-2 model would look like

\[
\begin{align*}
\pi_{0j} &= γ_0 + γ_0' Z_j + θ_j \\
\pi_{1j} &= γ_1' + γ_1' Z_j \\
&\vdots \\
\pi_{(G-1)j} &= γ_{(G-1)}' + γ_{(G-1)}' Z_j \\
\pi_{Gj} &= δ_i' \\
&\vdots \\
\pi_{(M-1)j} &= δ_i'_{(M-G)}.
\end{align*}
\]
Paper 2: A Reparameterized Rasch Model

Note that by incorporating covariates in the RRM the item difficulty estimates are no longer expected to be the same as those produced by the traditionally parameterized Rasch model.

**Multidimensional Random Coefficients Multinomial Logit Model**

The RRM assumes that responses associated with the mutually exclusive item groups do not depend on different latent traits. When this assumption does not hold, the data exhibit between-item dimensionality (Wang, Wilson, & Adams, 1997) and a multidimensional random coefficients multinomial logit model (MRCMLM) may be more appropriate (Adams, Wilson, & Wang, 1997). The MRCMLM treats each item group as if it were governed by its own latent trait, and the traits are allowed to correlate with one another. The difference between the RRM and MRMCMLM can be illustrated by the following level-2 model:

\[
\begin{align*}
\pi_{0j} &= \gamma_0 + \theta_{0j} \\
\pi_{1j} &= \gamma_1 + \theta_{1j} \\
& \vdots \\
\pi_{(G-1)j} &= \gamma_{(G-1)} + \theta_{(G-1)j} \\
\pi_{Gj} &= \delta_G \\
& \vdots \\
\pi_{(M-1)j} &= \delta_M.
\end{align*}
\]  

Unlike the RRM, each item group now has its own ability and the MRCMLM assumes that the abilities are distributed as multivariate normal random variables (although other distributional assumptions may be made, see Adams, Wilson, and Wang, 1997). Effectively, the RRM could be seen as the MRCMLM where all dimensions (all \(\theta_g\) distributions for \(g = 0\) to \(G-1\)) correlate with each other perfectly.

Using the RRM when the MRCMLM is more appropriate can be problematic. First, unidimensionality is a fundamental assumption of the Rasch model (Smith & Miai, 1994). When it does not hold, it brings parameter estimates into question. Second, using the RRM with multidimensional data may yield higher Type I error rates for the \(\gamma_g\)'s because the standard errors associated with them may be too small and the degrees of freedom used in hypothesis tests may be too large (i.e., based on the number of item responses at level 1 rather than the number of respondents at level 2) (Raudenbush & Bryk, 2002, Chapter 5). Consequently, it is important that the MRCMLM be ruled out if
the RRM is used, and this can be done in two ways. First, the deviances of the two models can be compared. In order to do this, the models must be fit using a full maximum likelihood procedure (see Hox, 2002, p. 110). Then a likelihood ratio test comparing the deviances of the RRM and the MRCMLM can be conducted (Raudenbush & Bryk, 2002, pp. 60-61) and/or the model with the smaller information criterion, such as AIC, can be selected (Burnham & Anderson, 2002, p. 61). Second, an estimate of the disattenuated squared correlation of the dimensions, $\hat{\rho}^2(\theta_j, \theta_k)$, can be used to gauge how much variance two dimensions hold in common.

**An Example**

Seventy professionals who worked with nonprofit organizations were surveyed at a professional development workshop at Stanford University. Among the survey questions were six that probed the social policy perspective (SPP) of the professionals. Respondents were asked to endorse or reject statements using a 4-category Likert scale. The responses were later dichotomized, with 1 reflecting a perspective on social policies that is consistent with the political left and 0 reflecting a perspective that is not. The six items were divided into three statements about the political left and three parallel statements about the political right (see Figure 1).

<table>
<thead>
<tr>
<th>Left Group</th>
<th>Right Group (Reverse Scored)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>With respect to social policy, I am a liberal.</td>
</tr>
<tr>
<td>Item 4</td>
<td>I consider my views on social policy to be left of center.</td>
</tr>
<tr>
<td>Item 3</td>
<td>I tend to favor social policies promoted by Democrats.</td>
</tr>
</tbody>
</table>

Figure 1. Parallel items used to gauge social policy perspective (SPP)

To construct a scaled score, the dichotomized responses to the statements about the right were reversed (that is, a 1 was rescored as 0, and a 0 was rescored as a 1) and then scaled using a Rasch model. The decision to reverse the scores was based in part on the assumption that statements about the left would yield responses similar to reversed responses to parallel statements about the right. To test the validity of this assumption, a traditionally parameterized Rasch model and an RRM were fit.

However, the interpretation of the RRM only makes sense if it is a reasonable model for the data. As described above, another possible model is the MRCMLM, so this alternative model was fit first to judge whether RRM should be used. The MRCMLM was fit as a two-dimensional, between-item model in which the items were divided into left and right groups as described by the equations given in (14). The results suggest that the RRM should be preferred for two reasons. First, the deviance statistics for the models
Paper 2: A Reparameterized Rasch Model

(as estimated using full maximum likelihood) differed by only 3.3 (1196.5 for the RRM estimating 7 parameters versus 1193.2 for the MRCMLM estimating 9 parameters). A likelihood ratio test comparing the deviances of the two models confirms that the RRM is preferred ($\chi^2 = 3.306$, $df = 2$, $p = .190$) as does a comparison of AIC statistics computed from the deviances (the smaller, 1210.5, for the RRM versus 1121.2 for the MRCMLM). Second, the disattenuated R-squared for the two dimensions indicates that the dimensions share 88% of their variance (a correlation of .94). Given this, the use of the RRM is warranted.

Table 3 presents the results of the traditionally parameterized Rasch model. The estimate for $\delta_0$ was -0.115 logits, equivalent to an average item difficulty of 0.115 logits and a base rate of 47% of items scored as a 1. While the base rate was less than might be expected from a group of respondents who were on average politically middle-of-the-road (a rate of 50%) it was not statistically significantly different from 50% ($SE = 0.273$, $t = -0.421$, $df = 69$, $p = .674$). Note that the $\delta_i$’s, ranged from -0.670 to 0.674, indicating that it was more difficult for some items to earn a score of 1 than others (i.e., the items appear to reflect a range of social policy perspectives).

Table 3

<table>
<thead>
<tr>
<th>Group</th>
<th>Item</th>
<th>Parameter</th>
<th>Reference</th>
<th>Estimate (Logits)</th>
<th>SE</th>
<th>t value</th>
<th>df</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>Item 1</td>
<td>$\delta_1$</td>
<td>Yes</td>
<td>-0.115</td>
<td>0.273</td>
<td>-0.421</td>
<td>69</td>
<td>.674</td>
</tr>
<tr>
<td>Right</td>
<td>Item 2</td>
<td>$\delta_2$</td>
<td></td>
<td>-0.002</td>
<td>0.282</td>
<td>-0.007</td>
<td>414</td>
<td>.994</td>
</tr>
<tr>
<td>Left</td>
<td>Item 3</td>
<td>$\delta_3$</td>
<td></td>
<td>-0.670</td>
<td>0.286</td>
<td>-2.345</td>
<td>414</td>
<td>.020</td>
</tr>
<tr>
<td>Right</td>
<td>Item 4</td>
<td>$\delta_4$</td>
<td></td>
<td>-0.670</td>
<td>0.286</td>
<td>-2.345</td>
<td>414</td>
<td>.020</td>
</tr>
<tr>
<td>Right</td>
<td>Item 5</td>
<td>$\delta_5$</td>
<td></td>
<td>0.479</td>
<td>0.286</td>
<td>1.676</td>
<td>414</td>
<td>.094</td>
</tr>
<tr>
<td>Right</td>
<td>Item 6</td>
<td>$\delta_6$</td>
<td></td>
<td>0.674</td>
<td>0.288</td>
<td>2.338</td>
<td>414</td>
<td>.020</td>
</tr>
</tbody>
</table>

Table 4 presents estimates from an RRM. Because the data are balanced, the estimate of $\gamma_0$ from the RRM equals the estimate of $\delta_0$ from the prior model (i.e., the mean of the group means is equal to the grand mean). The estimate for $\gamma'_i$ was -0.384, so the log odds of scoring a 1 on statements about the left was 0.768 logits less than the log odds of scoring a 1 on reverse-scored statements about the right. In other words, respondents found it easier to reject the originally worded statements about the right than to endorse parallel statements about the left. This difference is statistically significant ($SE = 0.129$, $t = -2.982$, $df = 414$, $p = .003$) and provides evidence that the assumption regarding the equivalence of parallel left group and (reversed) right group items does not hold.
Table 4  
Estimates from an RRM

<table>
<thead>
<tr>
<th>Group</th>
<th>Item</th>
<th>Parameter</th>
<th>Reference</th>
<th>Estimate (Logits)</th>
<th>SE</th>
<th>t value</th>
<th>df</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>--</td>
<td>Intercept</td>
<td>δ₀</td>
<td></td>
<td>-0.115</td>
<td>0.273</td>
<td>-0.421</td>
<td>69</td>
<td>.674</td>
</tr>
<tr>
<td>--</td>
<td>Left Group</td>
<td>γ₁'</td>
<td></td>
<td>-0.384</td>
<td>0.129</td>
<td>-2.982</td>
<td>414</td>
<td>.003</td>
</tr>
<tr>
<td>--</td>
<td>Right Group (rev.)</td>
<td>γ₂</td>
<td>Yes</td>
<td>0.384</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Left</td>
<td>Item 1</td>
<td>δ₁'</td>
<td>Yes</td>
<td>0.573</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Right</td>
<td>Item 2</td>
<td>δ₂'</td>
<td>Yes</td>
<td>-0.386</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Left</td>
<td>Item 3</td>
<td>δ₃'</td>
<td></td>
<td>-0.286</td>
<td>0.254</td>
<td>-1.129</td>
<td>414</td>
<td>.260</td>
</tr>
<tr>
<td>Left</td>
<td>Item 4</td>
<td>δ₄'</td>
<td></td>
<td>-0.286</td>
<td>0.254</td>
<td>-1.129</td>
<td>414</td>
<td>.260</td>
</tr>
<tr>
<td>Right</td>
<td>Item 5</td>
<td>δ₅'</td>
<td></td>
<td>0.095</td>
<td>0.254</td>
<td>0.374</td>
<td>414</td>
<td>.708</td>
</tr>
<tr>
<td>Right</td>
<td>Item 6</td>
<td>δ₆'</td>
<td></td>
<td>0.291</td>
<td>0.256</td>
<td>1.136</td>
<td>414</td>
<td>.257</td>
</tr>
</tbody>
</table>

Table 5  
Item Difficulty Estimates from a Traditionally Parameterized Rasch Model and the RRM

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Traditionally parameterized Rasch model</th>
<th>Reparameterized Rasch model (RRM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>δ₁</td>
<td>-0.074</td>
<td>-0.074</td>
</tr>
<tr>
<td>δ₂</td>
<td>0.117</td>
<td>0.117</td>
</tr>
<tr>
<td>δ₃</td>
<td>0.785</td>
<td>0.785</td>
</tr>
<tr>
<td>δ₄</td>
<td>0.785</td>
<td>0.785</td>
</tr>
<tr>
<td>δ₅</td>
<td>-0.364</td>
<td>-0.364</td>
</tr>
<tr>
<td>δ₆</td>
<td>-0.559</td>
<td>-0.559</td>
</tr>
</tbody>
</table>

In Table 5 the item difficulty estimates produced by the RRM and the traditionally parameterized Rasch model are compared. As expected, they are identical up the level of precision in the table (they differ slightly at higher levels of precision). It is clear that the RRM is fitting a Rasch model, but the way in which the RRM presented information has utility for testing the assumption of item group equivalence.

Given that the responses to the left and (reversed) right items do not appear to be equivalent, it might be of interest to determine if this lack of equivalence holds for all respondents. To test this, a covariate indicating years of professional experience was added. The relationship between years of professional experience and the equivalence of item group means was tested with the extended RRM described by the equations given in (13). An effect-coded variable identifying respondents with six or more years of experience (30 professionals coded as 1) versus those without (40 professionals coded as -1) was entered as a main effect and an interaction. The main effect (γ₀₁) quantifies the difference in the base rate of endorsement between professionals with more and less
experience. The interaction ($\gamma'_{12}$) quantifies the extent to which the two groups of respondents tended to have similar mean item difficulties for the left and (reversed) right item groups. In other words, the interaction tests whether the observed lack of equivalence in item group means is of similar magnitude across the two respondent groups.

Table 6

<table>
<thead>
<tr>
<th>Group</th>
<th>Teacher</th>
<th>Parameter Reference</th>
<th>Estimate (Logits)</th>
<th>SE</th>
<th>t value</th>
<th>df</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>--</td>
<td>Intercept</td>
<td>$\gamma_0$</td>
<td>-0.003</td>
<td>0.285</td>
<td>-0.012</td>
<td>68</td>
<td>.991</td>
</tr>
<tr>
<td>--</td>
<td>More Experience (main)</td>
<td>$\gamma_{01}$</td>
<td>0.571</td>
<td>0.285</td>
<td>2.004</td>
<td>68</td>
<td>.049</td>
</tr>
<tr>
<td>--</td>
<td>Less Experience (main)</td>
<td>--</td>
<td>-0.571</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>--</td>
<td>Left Group</td>
<td>$\gamma'_{11}$</td>
<td>-0.484</td>
<td>0.140</td>
<td>-3.469</td>
<td>412</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>and More Experienced (int.)</td>
<td>--</td>
<td>-0.400</td>
<td>0.140</td>
<td>-2.861</td>
<td>412</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>and Less Experienced (int.)</td>
<td>--</td>
<td>0.400</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>--</td>
<td>Right Group</td>
<td>$\gamma'_{2}$</td>
<td>0.484</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>and More Experienced (int.)</td>
<td>--</td>
<td>0.400</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>and Less Experienced (int.)</td>
<td>--</td>
<td>-0.400</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Left</td>
<td>Item 1</td>
<td>$\delta'_{1}$</td>
<td>0.572</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Right</td>
<td>Item 2</td>
<td>$\delta'_{2}$</td>
<td>-0.426</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Left</td>
<td>Item 3</td>
<td>$\delta'_{3}$</td>
<td>-0.286</td>
<td>0.254</td>
<td>-1.127</td>
<td>412</td>
<td>.261</td>
</tr>
<tr>
<td>Left</td>
<td>Item 4</td>
<td>$\delta'_{4}$</td>
<td>-0.286</td>
<td>0.254</td>
<td>-1.127</td>
<td>412</td>
<td>.261</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td>0.106</td>
<td>0.267</td>
<td>0.398</td>
<td>412</td>
<td>.691</td>
</tr>
<tr>
<td>Right</td>
<td>Item 5</td>
<td>$\delta'_{5}$</td>
<td>0.320</td>
<td>0.268</td>
<td>1.194</td>
<td>412</td>
<td>.234</td>
</tr>
</tbody>
</table>

To be certain that the RRM was appropriate it was again compared to a MRCMLM that incorporated the same main effect and interaction for experience level. The deviance was 1181.6 for the RRM estimating 9 parameters and 1180.4 for the MRCMLM estimating 11 parameters; a likelihood ratio test comparing the deviances of the two models was not statistically significant ($\chi^2$ = 1.215, $df$ = 2, $p$ > .500); and the RRM had the smaller AIC (1199.6 versus 1202.4). Again, the RRM is warranted.

Table 6 presents the results of the RRM with the covariate. The main effect for experience was positive (0.571 logits) and statistically significant ($SE = 0.285$, $t = 2.004$, $df = 68$, $p = .049$). This indicates that professionals with six or more years of experience tended to have social policy perspectives that were more aligned with the political left than their less experience peers. In fact, the odds that more experienced professionals would realize a score of 1 on an item were, on average, roughly 3 times the odds associated with less experienced professionals.

The experience-by-left-group interaction was negative (-0.400) and statistically significantly different from 0 ($SE = 0.140$, $t = -2.861$, $df = 412$, $p = .005$). This indicates that professionals with more experience were less likely to achieve a score of 1 on left items than they were on (reversed) right items. Put another way, the difference between the average difficulty of left and (reversed) right items was greater for professionals with
more experience, reflecting a greater tendency for more experienced professionals to reject (originally worded) right items.

Table 7
*Average Item Difficulty by Experience Level and Item Group*

<table>
<thead>
<tr>
<th>Item Group</th>
<th>Left</th>
<th>Right (Reversed)</th>
<th>Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Six or More Years of Experience</td>
<td>0.317</td>
<td>-1.452</td>
<td>1.768</td>
</tr>
<tr>
<td>Less than Six Years of Experience</td>
<td>0.659</td>
<td>0.490</td>
<td>0.169</td>
</tr>
</tbody>
</table>

The interactions are reflected in Table 7, which presents the average item difficulty by item group and experience level. For professionals with less than six years of experience, the left group was more difficult than the (reversed) right group by 0.169 logits on average. For professionals with six or more years of experience, the left group was more difficult than the (reversed) right group by 1.768 logits. The difference between these differences (i.e., the difference in the slopes) is 1.6 logits, which is a function of the four interaction estimates in Table 6. Figure 2 presents the interaction graphically and illustrates the magnitude of the estimated effect. Given the size of the interaction, it appears that the validity of the assumption that left items and (reversed) right items should produce similar scores may be associated with the level of professional experience of respondents.

Figure 2. The average difficulty of left and (reversed) right items for more experienced (six or more years) and less experienced professionals
Conclusion

A reparameterized Rasch model was presented and its benefits—the estimation of item group difficulties and associated tests of significance—were described. It was argued that the RRM provides one way of exploiting the structure of a measurement instrument to strengthen and improve evaluations. This paper illustrated how the RRM helped ascertain the validity of an assumption about reverse-scored items and the extent to which that assumption was associated with a characteristic of respondents (years of professional experience). In support of these purposes, the RRM was fit as a HGLM because of the flexibility it offers and the availability of software.

The example is a simple one intended to illustrate the application of the RRM, and some caveats should be offered. First, the MRCMLM ideally requires a larger number of cases and items than were found in the example. This is one explanation as to why the MRCMLM model did not fit well, notwithstanding the high estimated correlation between dimensions. Second, the example offers only one illustration of the RRM. The utility of RRM in experimental settings and other testing situations was not addressed, and future research in these contexts is needed. Third, theoretical concerns may sometimes outweigh technical criteria for modeling data. For example, if one had a strong basis for considering left and right items as representing distinct dimensions, one might choose to use the MRCMLM because of what it can reveal about the research question being investigated.

Nonetheless, the RRM would seem to provide researchers with a simple method of fitting data in a way that conforms to a Rasch model while simultaneously gathering information that directly relates to questions of instrument validity, population homogeneity, and program performance. The flexibility of the HGLM, which allows for the incorporation of covariates at the item and person levels, helps make the RRM a practical tool for researchers. For evaluators, who often work with limited resources, the ability of the RRM to simultaneously answer substantive questions and gather validity evidence would also seem to have practical appeal.
Paper 2: A Reparameterized Rasch Model

References
External Validity, Evaluation Evidence, and Action in Local Contexts

Introduction

Imagine that you were asked to choose between two crystal balls. The first can unerringly tell you whether a program improved the lives of past participants, and the second can unerringly tell you whether a program will improve the lives of future participants. If you are like most in our profession, the second is far more desirable because evaluation is action oriented—we engage the world in order to improve it. Descriptions of the past, whether to promote accountability or inform historical debate, would largely become irrelevant if we knew for certain the future consequences of policies and programs. Yet in spite of our desire to be forward looking, our profession seems fixated on the past. In particular, we have spent an inordinate amount of time arguing, often with raised voices, about whether evaluations with experimental designs are the best way to describe the past performance of a program, all the while neglecting the ways in which we can shed more light on future performance using the methodologically diverse evaluations we actually produce.

The main purpose of this paper is to shift the discussion of validity as it relates to outcome evaluation from the past to the future, from concepts of internal validity to those of external validity, and from the quality of evidence we can produce to how we use the evidence we have to make predictions and take effective action. I begin by describing the recent debates about credible evidence that have dominated the professional landscape. Then I provide a brief overview of the traditional validity framework as it relates to external validity and the longstanding search for alternatives. I discuss the applicability of the traditional framework to evaluation, examine three alternative strategies for strengthening predictions of future program performance based on past performance, and argue that only one, use strategies, has the potential to substantially improve evaluation practice in the near term. I then conclude by suggesting how stakeholders can use evaluation evidence to make predictions that inform action in local contexts.

What Works? What Do You Mean?

Over the past ten years, policymakers on both sides of the aisle have come to believe that evaluations can and should answer a now ubiquitous question—What works? Bipartisan faith in our ability to answer the question is exemplified by the education policies of Presidents George W. Bush and Barack Obama. As different as these two leaders may be, both initiated large-scale policies—No Child Left Behind and Race to the Top, respectively—in which they used evaluation to help their administrations fund “what works” (Bush, 2001; Obama, 2009) instead of “fancy theories” (Bush 2002), “what sounds good” (Bush, 2004), or “ideology” (Obama, 2008). The Bush administration was so confident that evaluation would produce credible evidence of effective programs that the U.S. Department of Education's Institute of Education Sciences created the What Works Clearinghouse to catalog and disseminate evaluation results (Institute of Education Sciences, n.d.). As a profession, this is exactly the sort of role that we have long argued evaluation should play (Campbell, 1988), and the fact that it has been endorsed by Republicans and Democrats would seem to signal that the profession may have arrived.

Yet it raises the issue of how evaluators can best answer the question “What works?” The question does not explicitly ask for a description of the past performance of a program (What worked?) or a prediction of future performance (What will work?). Rather it requires that
the past and the future be addressed simultaneously, the former with evaluation results and the latter with a line of reasoning that connects past results to future action. Using the past to inform the future is no simple matter, but it is made far more problematic because the question (What works?) suggests a false line of reasoning—to know past performance is to know future performance. This is what we typically mean when we assert that something works. If one turns the ignition key in an automobile and it starts, one may accurately state that the car works. By this we understand that it works today, it will almost certainly work tomorrow and the day after that, and it will work under a wide variety of conditions. A similar assertion cannot be made about programs because they famously do not perform consistently over time and across contexts. Thus the answer to the question “What works?” must always be a qualified answer, stipulating for whom, where, why, for what, and when the program will be effective. The challenge facing evaluators is how to use this qualified answer to construct a line of reasoning that produces reliable predictions about future program performance in specific contexts.

This is a formidable challenge; we lack a systematic history of concern with making predictions of this sort, and no theory is available for researchers who want to ensure that their evaluations better inform future action (Shadish, Cook, & Campbell, 2002, p. 342). This has not deterred our eagerness to proclaim and debate what does and does not work. Google Scholar identifies 1,720 pieces of academic writing with the phrase what works in the title that were produced over the past ten years in the social sciences, arts, and humanities. In medicine, pharmacology, and veterinary science, that number is only 153. This tenfold discrepancy should give us pause. It may merely reflect the popularity of an effective bit of political rhetoric—Who can argue against doing what works? However, it may reflect a willingness to draw too strong conclusions from weak lines of reasoning.

Recent debates about what constitutes credible evidence have shed new light on how lines of reasoning connecting past results to future performance can be strengthened. The underlying premise of the debate is not controversial—the more credible the evidence produced by an evaluation the more sound the conclusions based upon it. However, there is currently a fundamental disagreement in the field about what counts as credible evidence (see Donaldson, Christie, & Mark, 2008). While this dispute has an eerie similarity to the qualitative-quantitative “Paradigm War” of the 1970s and 1980s, the current “Causal War” is different in that it focuses largely on what constitutes scientifically-based evaluation methods (Scriven, 2008). A central theme in the debate is whether or not randomized control trails (RCTs) are the best design—the so-called gold standard—or even a good design for producing credible evidence about program effects, and whether that evidence provides an answer to the question “What works?”

Some evaluators argue that the problem is not whether evidence is credible, per se, but whether the evidence can be acted upon. In particular, by organizing evaluation resources around one small question—Does the program work?—they contend that evaluations can produce only one small answer (Greene, 2008). This answer may have some utility, but it is probably insufficient to guide the actions of policymakers at a national level, and it entirely misses the mark for the immediate stakeholders of a program. The people who work directly with programs or are directly affected by them want to know if a program will be effective in the local context in which they have the power to act.
Julnes and Rog (2008) noted that in order for evidence to be useful, it needs to be actionable as well as credible. In particular, it needs to be deemed appropriate for guiding actions in specific local contexts, and this is the motivating consideration for this paper. In an effort to shift the debate from credibility to action, I will make use of three concepts that require some clarification. Validity, the principal concept under discussion, is both an argument and a warrant. Evaluators must make the case that the evaluation evidence they produce can be used in a particular way for a particular purpose; this is the argument for validity. When others use evaluation evidence in a given manner, they justify their action by citing a relevant validity argument that they accept; this is validity used as a warrant. Of course, evaluators may argue for one use and others may put evaluation evidence to different uses, which places the burden of making the argument upon the latter group. From this perspective, validity is not an on-or-off quality of the evaluation or a characteristic of evidence, but a responsibility shared by the evaluator and the user, a key point in our later discussion of alternatives.

There are many types of validity, and I am concerned with what has traditionally been called external validity and how it warrants prediction. If one accepts an argument for external validity made by an evaluator or a user of the evaluation, one feels justified in using the results of a past evaluation to make certain predictions about the future results of programs. However, each of us forms our own beliefs about the persuasiveness of the validity argument, and given our varied research traditions, tolerances for uncertainty, prior beliefs, personal stakes in programs, and evaluation approaches, consensus is the exception. In order to act on our predictions, we must somehow develop a shared faith that the results of past evaluations are believable; reflect the results of programs as a whole, not only the effects of interventions implemented by programs; and can predict the future results of other programs. Faith in this sense is arrived at individually, but faith can only be acted upon collectively, which brings us back to the point that validity is a shared responsibility. The extent to which we share it and with whom we share it are key concerns of modern evaluation practice that differentiate it from traditional research.

The Traditional Validity Framework: External Validity, Generalizability and Prediction

Describing the traditional validity framework and how it relates to external validity is challenging because the framework continues to evolve after almost a century of development. As early as 1923, McCall’s *How to Experiment in Education* offered a comprehensive framework, albeit without using the word validity. It cataloged a series of experimental designs that could be used to determine the effectiveness of educational interventions, each of which was represented by letters separated by dashes, for example *IT-EF-FT* for initial test (pretest), experimental factor (treatment), and final test (posttest), that foreshadowed the *X-O* diagrams so strongly associated with Campbell. McCall argued that the value of experimentation was that it should lead to the widespread use of more effective and efficient practices, but did not explain how experimental results could be constructed or used in ways that would accomplish this, focusing instead on the theoretical and procedural aspects of how to conduct experiments. By the time Greenwood (1945) and Chapin (1947) presented their own systematic approaches to using experiments to evaluate social interventions, they recognized that focusing on designing a good experiment often detracted from findings that could be applied widely, what Greenwood (p. 93) called trading significance for accuracy. Again, they offered few suggestions about how to design experiments that facilitated widespread action based on experimental results beyond...
Chapin’s (p. 187) suggestion that researchers should engage in a program of experimentation in which they purposefully replicated studies (something that he acknowledged was a rarity).

One of the great contributions that Campbell made was that he reorganized and expanded upon this nascent body of work, creating a more coherent, correct, and useful framework. At its heart is the distinction between internal validity and external validity (Campbell, 1957). The former provides a basis for believing a study’s result, which takes the form of a quantitative estimate of the actual treatment-control contrast given all of the particulars of a study’s circumstances. The latter provides a reason to believe that a study’s result is generalizable, which is to say that one can expect a similar result if the study is replicated under similar circumstances. I use the term circumstances as a catchall for everything that could be different between two replications of a study, in particular persons, settings, treatments, outcomes, and times, which are the five dimensions that Campbell and his peers most explicitly addressed (Cronbach, 1982; Cook, 1993).

Within this framework, generalizability depends on whether studies are conducted under similar circumstances. The similarity of two studies is ensured (over the long run, at least) if their respective circumstances are representative of a more general class of circumstances, and this can only be ensured (over the long run) if the circumstances are randomly sampled from a well-defined population of possible circumstances. This chain of reasoning is long, and when it is applied to even the most well understood case, using a random sample of participants to generalize to the population from which it was drawn, it almost always falls apart. Researchers typically start with a pool of volunteers from which they may sample a smaller group of participants. Some of those selected will decline to participate in the study, discontinue their participation during the course of the study, or participate yet refuse or be unable to provide data to researchers, netting a study sample that no longer represents those who volunteered and, more importantly, is unlikely to represent those who did not. Sampling from settings, treatments, outcomes, and times presents other problems, such as the impossibility of sampling the past or being able to identify and sample from a population of treatments.

Campbell conceded that these problems were largely insurmountable and concluded that random sampling was rarely a practical means of ensuring generalizability (see Campbell, 1986, p. 71). This led him to suggest an alternative logic for generalization that was based on what he called the principle of proximal similarity—that scientists generalize with most confidence to applications most similar to the setting of the original research. He acknowledged that this was something of a punt, a “metatheoretical basis for justifying a seemingly atheoretical rationale and approach to the generalization of findings” (p. 73). The irony is that while his is considered by some to be an unwavering proponent of quantitative methods, he argued for a wholly qualitative approach to generalization in which, “the principle of proximal similarity is normally (and it should be) implemented on the basis of expert intuition” (p. 76).

Throughout, Campbell de-emphasized concepts related to external validity in favor of those related to internal validity, which for him was “the prior and indispensable consideration” forming the “basic minimum” for experimental research (1957, pp. 310, 297). That is not to say he did not believe external validity was important, but rather that he could not fathom why one would want to generalize a result one did not believe. Moreover, he, like Greenwood before him, worried that researchers attempting to strengthen the generalizability of their work might be undermining their own efforts because internal and external validity “are to some extent
incompatible, in that the controls required for internal validity often tend to jeopardize representativeness” (Campbell, 1957, p. 297). Campbell never fully resolved the tension he created by characterizing internal validity as both a prerequisite for and detriment to external validity (Jimenez-Buedo & Miller, 2009), but instead advised researchers to prioritize internal validity and, if necessary, sacrifice external validity—advice that some contemporary evaluators engaged in the Causal War may have taken too much to heart.

The crisp distinction promised by the labels internal and external never fully materialized because they do not adequately distinguish the inferences researchers make from the object of experimental investigation. From Campbell’s perspective, experimental investigation sheds light on the actual treatment-control contrast. An experiment loses its utility if the actual contrast differs from the contrast of interest to researchers. Under Campbell’s original validity framework, the utility of the experimental contrast was not directly addressed. It was therefore possible to implement an experiment with strong internal and external validity that had little utility.

To see how this could happen and how Campbell addressed the problem, we can look to an example that he used in 1957 and returned to in 1986—placebos. Knowledge that one is being treated (a placebo effect) can influence experimental outcomes, but it usually is not a part of the potential effect that is of interest to drug researchers (the pharmacological effect). To test the pharmacological effect of a drug rather than a mixed pharmacological-placebo effect, Campbell suggested implementing an experiment with multiple treatment-control contrasts. This would, in Campbell’s words, “purify” the “total X stimulus complex” (1957, pp. 309-310). Following his advice, a researcher might implement an experiment with a nothing control, a placebo-only control, and the pill-based treatment under investigation. From these multiple contrasts the experimenter would be able obtain an estimate of the pharmacological effect. On the other hand, a simpler experiment that included only one contrast—a nothing control versus the pill-based treatment—would confound the pharmacological effect with a possible placebo effect.

The confusing point is that under Campbell’s original framework, the one-contrast and multiple-contrast experiments have equally strong internal and external validity, holding all else constant. That is, both can describe well the result of the experiment and the circumstances under which a future experiment would yield a similar result. However, the point of the experiment is to help medical practitioners discover dependable pharmacological effects not dependable experimental procedures. The multiple-contrast experiment is clearly more useful for this purpose.

Campbell (1986) recognized this weakness in the original framework, so he endeavored along with Cook (Cook & Campbell, 1979) to introduce the concept of construct validity (sometimes referred to as the construct validity of treatments or the construct validity of causes). Construct validity involves making inference from the particulars of a study to the higher-order constructs they represent (Shadish et al., 2002, p. 65). The causal construct of interest is an intended intervention, and construct validity depends on the similarity of the actual treatment-control contrast to the intervention. In the drug example, a researcher may want to use experimental evidence about the efficacy of a specific dosage of chemicals ingested in pill form to draw conclusions about the efficacy of the pills. The alignment in this case is quite high. In education, a researcher may want to use experimental evidence about the effectiveness of a
tutoring program to draw conclusions about the effectiveness of a tutoring policy or the effectiveness of a particular approach to tutoring. While the construct and the contrast are arguable still aligned, they are less aligned than they were in the drug example.

In both cases, the researcher is making inferences as scientists would—from the particular to the general regarding an intervention. However, interventions are only one action undertaken by the larger organizations that implement them (I refer to these organizations as programs). Organizational issues such as labor relations, cash flow, and competition have the potential to impact the results of any intervention the organization implements, but these issues are rarely illuminated by experiments. Consequently, stakeholders not only find themselves making different inferences than scientists—from one particular context (an experiment) to another particular context (an individual, local program)—they must do so with a weaker evidentiary basis.

**A Longstanding Search for Alternatives**

As the traditional framework evolved over the years, so did criticisms offered by a number of prominent evaluators who questioned the applicability of the traditional framework to evaluation as they practiced it. Stake (1978) hoped to replace, or at least augment, the traditional concept of generalization with what he called *naturalistic generalization*, which he argued is “arrived at by recognizing the similarities of objects and issues in and out of context and sensing the natural covariation of happenings” (p. 6). According to Stake, unlike the traditional concept of generalization, naturalistic generalization is born of intuition and tacit knowledge of the world, promotes a thorough understanding of the particular, is action oriented, and avoids misunderstandings and simplistic thinking. Importantly, he also held that it shifts the burden of responsibility for generalization from the evaluator to users of evaluations.

Similarly, Guba and Lincoln (1989) introduced the concept of *transferability* because unlike traditional generalizations, which depend on a level of representativeness that can rarely be assured in practice, transferability depends on the capacity of stakeholders to compare programs previously evaluated to a future action. Thus, by determining which salient conditions overlap or match between the past and future implementations, the transferability of the program can be judged. Evaluators can promote transferability, they argued, by providing thick descriptions of programs, their contexts, and results. With this in hand, stakeholders engage in a particular form of collaborative group process (a dialectic) that Guba and Lincoln asserted would help stakeholders arrive at better judgments about transferability. Thus, as with Stake’s naturalistic generalization, transferability shifts the burden of proof for generalization to the receiver (i.e., the user of an evaluation).

Cronbach (1982) supplied perhaps the strongest criticism of the traditional framework, at least as it was imagined by Campbell, by a researcher firmly rooted in the orthodoxy of quantitative research. Among other things, he introduced the notion of *reproducibility* and described it as taking place at three levels: (1) exact replications of an intervention using complete information about how to implement it, (2) close replications of an intervention using the best available information about how to implement it, and (3) rough approximations of an intervention when no information about how to implement it is available. To judge reproducibility, stakeholders perform thought experiments in which they consider a historical event that can never be repeated (the past implementation of a program) and how it might unfold under different circumstances. Imagination is the key, and Cronbach described the cognitive
process that allowed stakeholders to harness their imaginations to arrive at inferences that could be discussed, debated, and acted upon.

None of these suggested alternatives to traditional concepts of external validity and generalizability took root, and most practicing evaluators are probably unaware of them. However, suggestions such as these reflect a persistent sentiment within the field that the traditional framework does not fully serve evaluation as it is practiced. Some criticisms stem from fundamental differences in how evaluators believe research should be conducted, and the evaluators who make them are unlikely to be satisfied by anything short of revolutionary change. On the other hand, as exemplified by Cronbach, there are evaluators with more conventional views regarding research methods who have also perceived a misfit. It is from this perspective that this paper explores how evaluators can more fully reconcile the traditional framework with modern evaluation practice.

**Applicability of the Traditional Framework to Modern Evaluation Practice**

The traditional framework has been a workhorse for evaluators for over 50 years. While it is fundamentally sound, it has a number of features that limit its applicability to modern evaluation practice. In particular, I discuss five features. First, the traditional framework is concerned with generalizability—that one can expect a similar result under similar circumstances—when what is more relevant is predictability—that one knows what result to expect under a variety of circumstances, some of which we may have never encountered before. Dissimilar circumstances are not necessarily inferior circumstances. If a program is replicated with a longer intervention we would not expect the same results, we would likely expect better results. The entire enterprise of program improvement is based upon our ability to imagine untried circumstances that will be more conducive to achieving a desired result, while the traditional approach to generalization denies that we can do so with any reliability.

Second, the traditional framework was intended to help researchers generalize the results of experimental research, not help stakeholders predict the performance of programs in specific, local contexts. The traditional framework organizes the strategies that scientists use to make generalizations with and without the benefit of a formal sampling scheme (Shadish et al., 2002, Chapter 11). While the logic underlying these generalizations is sound, scientists and stakeholders use different types of information to understand different constructs. Scientists, in the Campbellian tradition, use experimental results to generalize about interventions. Stakeholders use results from studies designed in many ways, most of which are not experimental, to predict the future performance of programs. Programs are larger in scope and far messier than the treatment-control contrasts studied by scientists. Consequently, application of the traditional framework to stakeholder settings raises questions related to construct validity.

Third, dependence on expert intuition is troubling. It can easily be construed as taking control out of the hands of stakeholders and placing it into those of powerful elites who are not directly impacted by the program. One can avoid this by defining expert as stakeholder, and Campbell may well have endorsed this definition, but the more serious problem is that regardless of who is in control, characterizing the judgment of similarity as intuition implies that it is unconscious, mystical, and opaque. What would be preferable is a process that is intentional, logical, and transparent. This not merely championing a platitude; there is evidence that an informal yet systematic approaches with these qualities can improve predictions, even when they
lack some of the characteristics found in more formal algorithms and scientific thinking (Dawes, 1979; Dawes, Faust, & Meehl, 1989; Gigerenzer & Goldstein, 1996; Meehl, 1954).

Fourth, if the purpose of an evaluation is to answer some variation of the question “What works?” it is important to remember that what works depends on what matters. The values that lead one person, group, or organization to declare success may lead others to declare failure or shrug their shoulders with uncertainty. This may reflect a normative mismatch—reasonable people disagreeing about what programs should strive to achieve. Or it may reflect an empirical mismatch—reasonable people disagreeing about how to measure the extent to which a commonly agreed upon end was achieved. Differences of this sort are a healthy part of the messy democratic process, yet they do not fit easily into the tidy confines of the traditional framework developed by experimenters who are far more likeminded than the stakeholders, policymakers, and general public they serve.

Fifth, the traditional framework assumes that programs are static, well documented, replicable entities. That is, programs are equivalent to the protocols that experimenters use to conduct and replicate their research. The reality is that programs are only as well developed as they need to be to be funded, which is far from the level of detail that would allow them to be implemented with consistency. Thus, any particular implementation of a program is an unstable amalgam of intention, improvisation, responsive local adaptation, ongoing improvement efforts, funding requirements, and staff competence. From one implementation to the next, these factors can vary substantially. While this might appear to be sloppiness, it is often the result of diligence. More and more, programs find that they must operate according to a fee for service model in which they compete by customizing their programs for every client or request for proposals. At the same time, clients and more traditional funders learn from their experience and from research, regularly changing their requirements to take advantage of what they learned. And it is the rare program that does not reflect on its prior implementations and make improvements for the next. Given this reality, generalizations or predictions about programs will only be relevant if they take into account not only random treatment variation, but variation that is intentional, continuous, and likely beneficial.

Three Broad Strategies for Improving Prediction

Design Strategies

From the traditional perspective, evaluation evidence should, whenever possible, be gathered using evaluation designs that support generalization. As the framework evolved, the uses subsumed under the label of generalization expanded. In particular, Cook (1993) and later Shadish et al. (2002) discussed generalization, interpolation, and extrapolation, which collectively get closer to what I mean by prediction, and suggested five principles that evaluators can exploit in order to design experiments and quasi-experiments that support these inferences. In line with Campbell’s perspective, these authors argued that generalizing from a source (the past) to a target (the future) is logically warranted when the populations, treatments, outcomes, settings and times of the source and target are:

1. very similar (Principle 1: Surface Similarity);
2. different but not in ways that influence the effectiveness of the program (Principle 2: Ruling Out Irrelevancies);
different, influence the effectiveness of the program in undesirable ways, but are identified and can be avoided or controlled (Principle 3: Making Discriminations);

(4) manipulatable in such a way that program effectiveness can be controlled or predicted as it can be when there is a known dose-response relationship (Principle 4: Interpolation and Extrapolation); or

(5) described by a well crafted causal theory that can be used to explicate results (Principle 5: Causal Explanation).

Evaluators can discover how strongly these principles hold by designing their evaluations in ways that allow them to analyze variation in results. That variation can be naturally occurring (men versus women) or purposefully introduced (ensuring a range of ages or using more than one version of the treatment); it can be explored in a single study setting (one program) or across many settings (multiple programs); and it can be studied in multiple locations simultaneously (multi-site evaluations) or sequentially (replication). In short, if one can estimate how variation in circumstances (populations, treatments, etc.) relates to variation in results, one has a logically sound basis for controlling or predicting future results across (and at times beyond) the range of circumstances considered.

These design strategies are conceptually sound, and they can be applied to studies that look nothing like an experiment. Five replications of a naturalistic evaluation provide a better basis for prediction than one (assuming that you have faith in the evaluation results and that they represent program results). Moreover, these design strategies provide better answers to the questions stakeholders have because they describe for whom, where, why, for what, and when the program is effective.

Why, then, do these strategies fail to satisfy? The problem is not logic but logistics. Virtually every program that receives government, philanthropic, or corporate support must conduct an evaluation. Among them are the approximately one million registered nonprofit organizations in the United States (National Center for Charitable Statistics, 2009). Reed and Morariu (2010) surveyed a representative sample of nonprofits and found that 85%—roughly 830,000—evaluated some part of their work over the past year. This estimate does not reflect the numerous public schools, government agencies, municipalities, military branches, for profit companies, and other organizations that conduct evaluations, nor do they take into account that many organizations perform multiple evaluations each year. Thus, the number of evaluations conducted yearly is easily measured in the millions.

The majority of evaluations for nonprofit organizations are not conducted by evaluation professionals. Dedicated evaluation staff or external evaluators are responsible for evaluations in only 21% of nonprofit organizations; non-evaluation staff (62%) and no one (17%) are responsible in the remainder. Evaluations in the nonprofit sector are used for multiple purposes, but the two leading uses are planning and revising programs (79% of organizations) and reporting to funders (74%). Consequently, at least in the nonprofit sector, evaluations are used primarily to make predictions about the future performance of programs that impact funding, expansion, and improvement. Also according to Reed and Morariu, nonprofit organizations spend very little on evaluation. Of the organizations that made expenditures on evaluation, 73% allocated 5% or less of the organization’s budget to that purpose. This is perhaps understandable
given that 36% reported that none of their funders supported their evaluation work. In fact, 84% stated that insufficient funding was a challenge that limited their evaluation efforts.

In the nonprofit sector, the best evidence we have is that the number of evaluations conducted is high, the expertise of evaluators is low, the level of funding for evaluation is modest, and evaluations are used most of the time to predict the future performance of programs. In other sectors where funding and evaluation expertise are not found in abundance, like public education and government, evaluation may be on a similar footing. Accordingly, design strategies like those advocated by Shadish et al. (2002)—as desirable as they may be—do not constitute a general solution because they require studies that are larger, more expensive, and more complicated than we can expect. This situation is unlikely to change without a fundamental shift in the way evaluations are funded and staffed. For the foreseeable future, we need to make the best of what we have.

**Reporting Strategies**

A strong push for better and more comprehensive reporting has come from many quarters, especially from those who want to develop a corpus of evidence-based programs. There are so many reporting standards currently being promoted that it would seem impossible to meet them all. They include CONSORT (Consolidated Standards of Reporting Trials; Altman et al., 2001), TREND (Transparent Reporting of Evaluations with Nonexperimental Designs; Des Jarlais et al., 2004); Standards for Reporting on Empirical Social Science Research in AERA Publications (AERA, 2006); RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance; Glasgow, Vogt & Boles, 1999; Glasgow et al, 2006), the Evaluation Report Checklist (Miron, 2004); and the Program Evaluation Standards (Yarbrough, Shulha, Hopson, & Caruthers, 2011). In addition, evaluators must satisfy the reporting standards of funding agencies (this may simultaneously entail the competing requirements of state agencies, federal agencies, and third-party organizations evaluating the state and federal agencies), and the divergent standards of the peer-reviewed journals in which evaluators may publish.

Unfortunately, limitations of time and space will result in some important pieces of information going unreported, even by those who are guided by rigorous reporting standards. Moreover, consistently reporting information that is salient to stakeholders, in particular organizational strengths, weaknesses, and values, would mean creating additional reporting standards. Even if they were added, social science researchers and evaluators are, for the most part, not qualified to gather and interpret this sort of information. Standards can improve the quality of reports, but the variety of standards being promoted reflects the fact that there is a divergence of opinion about what constitutes credible evidence. All standards for all users cannot be met, and barring an unprecedented convergence of opinion among those in the social sciences, policy, and program arenas, users will continue to lament the quality of the evidence at their disposal.
Use Strategies

Consider a mythical creature—a perfectly designed and executed experiment with unassailable internal and external validity. This experiment finds that an intensive afterschool program substantially improved the reading of third-grade students. A large group of stakeholders, living in a setting identical to that in which the experiment was conducted, have read the study and are considering whether to implement the program. They ask themselves if they can afford it and if they can implement it with fidelity, and they answer, “Yes.” They ask themselves if they want third-grade students to read better, and they answer, “Yes.” And after a great deal of thoughtful deliberation they ask themselves if they want to implement the program and they answer, “No.” Did they use the research evidence wisely?

The answer to that question depends on the logic underlying the decision, which depends to a large degree on the stakeholders’ program theory. Making predictions about the future effectiveness of a program based on past evaluations is like peeking through a keyhole in order to learn what is on the other side of the door. One gets a very clear view of some things while others are completely obscured. If the stakeholders’ program theory led them to suspect that the intensive nature of the program would cause students to spend less time studying other subjects and consequently learn less about them, it can be argued that stakeholders used the available evidence wisely—even though the experiment provided no evidence one way or the other about this potential side effect.

When we make predictions about the future performance of programs, we consider everything that we care about, not everything for which there is evidence or the quality of the evidence is high. I have argued that it is not feasible to improve predictions, at least in the foreseeable future, by consistently applying design strategies on a wide scale, even though at a conceptual level these strategies are appealing and should be used whenever possible. And while more comprehensive reporting strategies may be advantageous in many ways, some of the credible evidence produced by an evaluation will inevitably be left on the cutting room floor. So we must learn to live with the fact that the quality of the evidence at our disposal will always be weaker than evaluators, stakeholders, and policymakers desire. Evidence is king, but if evidence is thin program theories must suffice.

Our reliance on program theory can be illustrated further by mapping how the five principles for generalized causal inference presented by Shadish, et al. (2002) warrant the three types of claims made by those predicting the future performance of programs (Table 1). These claims can be abbreviated like begets like (similar programs and circumstances should produce similar results), different begets like (different programs and circumstances will in spite of their differences produce similar results), and different begets different (different programs and circumstances will produce different—presumably better—results). I do not consider the claim like begets different because, as the old chestnut warns, repeating the same action and expecting different results is one of the definitions of insanity.

The claim most closely associated with the traditional perspective on generalization, like begets like, can largely be warranted by some combination of the first two principles, surface similarities and ruling out irrelevancies. In concept, a line of reasoning supporting this claim can be constructed easily from evaluation evidence because one only needs to demonstrate that two sets of circumstances are reasonably similar. In practice, one is unlikely to find similar circumstances. The orthodox view, as Shadish et al. (p. xv) make clear, is that material
differences are the norm, thus “all causal statements are contingent.” Reliable predictions must therefore be based on more than similarity.

TABLE 1

<table>
<thead>
<tr>
<th>Five Principles of Generalized Causal Inference</th>
<th>Claims Made by Those Predicting the Future Performance of Programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principle 1: Surface Similarity</td>
<td>Like Begets Like</td>
</tr>
<tr>
<td>Principle 2: Ruling Out Irrelevancies</td>
<td>X</td>
</tr>
<tr>
<td>Principle 3: Making Discriminations</td>
<td>X</td>
</tr>
<tr>
<td>Principle 4: Interpolation and Extrapolation</td>
<td>X</td>
</tr>
<tr>
<td>Principle 5: Causal Explanation</td>
<td>X X</td>
</tr>
</tbody>
</table>

Unfortunately, when the circumstances surrounding a yet-to-be-implemented program differ substantially from those of programs that have been evaluated, the available empirical evidence becomes less relevant. Stakeholders who wish to judge whether a program implemented in different circumstances will produce a similar result are unlikely to find evidence that demonstrates that specific irrelevancies can be ruled out, discriminations made, or results interpolated or extrapolated. And a belief that different circumstances will give rise to different results—a cause-and-effect relationship not previously identified by research—can only be justified by program theories that have little or no empirical justification. When predicting the future results of programs implemented in different circumstances, stakeholders must rely heavily upon program theories that are a far cry from the well explicated, evidence-based theories that Shadish et al. describe. However, as imperfect as these theories may be, they have the potential to improve prediction if they allow stakeholders to incorporate evidence into their predictions in a more systematic way.

**Toward a More Systematic Process**

The best we can hope for, then, is to explicitly surface our theories about why a program should work and use them to guide predictions of whether it will work in a specific local context. Because external validity is a shared responsibility, these judgments should emerge from a group process in which the widest range of experts and non-experts, including direct and indirect stakeholders, work together. This coincides with Campbell’s (1993) contention that whatever
“objectivity” science may achieve, it is arrived at through shared plausibility judgments rather than proof. Building on this idea, I describe a systematic way of engaging stakeholders in a process of constructing shared plausibility judgments of the future effectiveness of programs. It is not the only way, and there is no empirical evidence to suggest that the particular solution offered here is the best, but as mentioned previously there is evidence that systematic solutions similar to the one described below can improve predictions.

The first step in the process is for stakeholders to surface a program theory. A great deal has been written about how to do this (Chen, 1990; Donaldson, 2007; Knowlton & Phillips, 2009) and how a program theory should reflect a shared understanding of which outcomes stakeholders hope to achieve and how the various characteristics of the program, participants, and setting are expected to interact to bring about those outcomes (see Donaldson, 2001, 2007; Donaldson & Gooler, 2003). With their program theory in hand, stakeholders review evaluations of candidate programs. The purpose of the review is to determine whether stakeholders share a faith in the believability of the evaluation results, the relevance of the results to the programs that were evaluated, and predictions based on the results. Believability is addressed when stakeholders answer the question, “Are some or all of the evaluation findings related to program effectiveness credible?” (See Appendix A for a data-collection form that scaffolds this process.) Relevance is addressed by answering the questions, “Do the evaluation results reasonably represent the results of the program that was evaluated?” and “Does the evaluation provide evidence about some or all of the outcomes in our program theory?” These deceptively simple questions may require that stakeholders engage in considerable discussion before they agree on answers.

If stakeholders answer “No” to any of the questions, the evaluation under consideration will not inform predictions and stakeholders should look to other evaluations for evidence that will serve their needs. If they answer “Yes” to all three questions, then stakeholders dig deeper, establishing the extent to which they believe past results predict future results. They do this by considering each outcome in their program theory, summarizing the evaluation evidence, indicating whether they consider it credible, and judging whether the level or quality of the observed outcome meets their definition of success. Then they consider for whom, where, when, and why the past program was implemented, describe the similarities and differences between these contextual factors and the ones they specified for their own program theory, and predict whether implementing the program in their local context would yield results that were better, the same, or worse than prior results. This is accompanied by a brief explanation of the logic underlying the predictions.

At the end of the process, if there is a consistent pattern of prior evaluation results meeting the current standard of success, and judgments that future circumstances will lead to the same or better results, the decision to implement the program becomes quite easy. More likely, the results will be mixed, leading to more discussion and debate, and possibly a decision to modify an existing program, combine elements of several, or design an entirely new program. The critical feature is that the final predictions, the actions that follow from them, and the logic underlying both, are well documented. If that is the case, it becomes possible to monitor programs as they are implemented and make midcourse corrections that increase the likelihood of success.
Conclusions

We care about validity because it reduces risky leaps of faith to more manageable hops of faith. In the end, however, it always comes down to faith—faith in the results of past evaluations, faith that those results represent the effectiveness of the program, and faith that past results are a reasonable basis for predicting future success. Yet the debate is not exclusively about how we come to have faith. In fact, I have worked hard to dodge the question of how we come to have faith in past evaluations results. That falls under the traditional heading of internal validity and it has consumed a disproportionate amount of our profession’s time and attention. Instead, I have focused on how we can use the faith we have to make predictions about the future effectiveness of programs in ways that can help us take effective action.

At the heart of my argument is the claim that research design and reporting standards, while useful, have a limited capacity to improve the quality of evaluation evidence on a wide scale. Consequently, meaningful improvements in how we act can only come from improvements in how we use the evidence at hand. I have suggested a systematic process for using evidence to make predictions that inform action, but the precise details of that process are not important. What is important is that the process be transparent, collective, and systematic. If this is case, not only may predictions improve, but we may be able to monitor and adjust programs more effectively over time. Even experts may get most of their predictions wrong, so we should strive, in the spirit of the statistician, to get them the least wrong, and in the spirit of the manager, to make frequent mid-course corrections to increase our chance of success.

Throughout the process for making predictions, stakeholders should play a central role. Without the inclusion and leadership of stakeholders, the values and concerns of those closest to programs will not be adequately addressed. The sort of comprehensive, local deliberation suggested here may be difficult to attain in practice because it requires time and resources that may be in short supply. Yet this process is worth promoting because the adage, “We predict the future from the past at our own peril,” is not entirely accurate. When we evaluators make predictions about programs and policies, it is others, often the most vulnerable, who are at peril and suffer if we are wrong. If we cannot consistently depend on evidence of higher quality, we can at least be smarter about how we use the evidence we have. And we can put those who are most directly affected into the driver’s seat. Evaluations serve many different groups and they employ a variety of theoretical stances, practical approaches, and research methods. The profession should reflect this reality, and it is what motivated the suggestion of a group process that does not depend on endorsing any one evaluation ideology. However, this suggestion is not a condemnation of the traditional validity framework, but a testament to how fundamentally sound it is that it can support extensions and adaptations of this kind.
Appendix A: Scaffolded Questions for Evaluation Evidence Review

**Believability**

Are some or all of the evaluation findings related to program effectiveness credible? Yes/No

**Relevance**

Do the evaluation results reasonably represent the results of the program that was evaluated? Yes/No

Does the evaluation provide evidence about some or all of the outcomes in our program theory? Yes/No

**Prediction**

<table>
<thead>
<tr>
<th>Outcome 1</th>
<th>Outcome 2</th>
<th>Outcome 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary of findings: The findings ARE NOT/ARE credible This result DOES NOT/DOES meet our standard</td>
<td>Summary of findings: The findings ARE NOT/ARE credible This result DOES NOT/DOES meet our standard</td>
<td>And so on…</td>
</tr>
</tbody>
</table>

**Similarities**

- What is the plausible effect of similarities and differences on outcomes?

<table>
<thead>
<tr>
<th>Outcome 1</th>
<th>Outcome 2</th>
<th>Outcome 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>For whom? How do past participants compare to our participants?</td>
<td>WORSE/SAME/BETTER than prior result or CANNOT SAY Because...</td>
<td>WORSE/SAME/BETTER than prior result or CANNOT SAY Because...</td>
</tr>
<tr>
<td>Where and when? How does the past setting compare to our setting?</td>
<td>WORSE/SAME/BETTER than prior result or CANNOT SAY Because...</td>
<td>WORSE/SAME/BETTER than prior result or CANNOT SAY Because...</td>
</tr>
<tr>
<td>Why? How does the past program theory compare to our program theory?</td>
<td>WORSE/SAME/BETTER than prior result or CANNOT SAY Because...</td>
<td>WORSE/SAME/BETTER than prior result or CANNOT SAY Because...</td>
</tr>
</tbody>
</table>
References


