Integrating Fidelity Data into the Analysis of Outcomes: Statistical Methods for Reducing Bias

Journal:	*Journal for Research on Educational Effectiveness*
Manuscript ID:	Draft
Manuscript Type:	Methodological Studies
Keywords:	fidelity, measurement error, regression calibration, effect size

Integrating Fidelity Data into the Analysis of Outcomes:

Statistical Methods for Reducing Bias

Abstract

Data assessing the fidelity of implementation is routinely collected as a part of the quality control of educational interventions. This paper proposes methods for using these data in evaluation of the impact of the intervention, rather than merely as a quality control check. The idea is that the effect size of an intervention may be decreased by poor implementation, and without information about the size of this decrease, the theory underlying the intervention cannot be properly tested. The proposed methodology provides an estimator of the maximum possible effect size that could occur if a full implementation were possible. This method also accounts and corrects for imprecise measurement of fidelity that frequently occurs due to either measurement or sampling error, or both. The methodology is illustrated with data from a longitudinal randomized control trial of a reading intervention.

Keywords: fidelity, measurement error, regression calibration, effect size

Integrating Fidelity Data into the Analysis of Outcomes:

Statistical Methods for Reducing Bias

Fidelity of implementation (i.e. treatment fidelity, treatment integrity) in educational research means the extent to which theoretically meaningful components of an intervention are realized in practice; i.e., the extent to which an intervention is implemented as intended (Gall, Gall, & Borg, 2007). In educational research, careful attention has been given to the measurement of dependent variables, or outcomes, but much less to the measurement of the independent variable, or treatment. Yet, without measures of fidelity, one cannot determine whether an experiment that does not confirm an intervention's effect is due to poor implementation, or failure of the theory underlying it. Because of the recognition of the crucial role that fidelity plays in interpreting experimental outcomes, it has recently been a topic of research and discussion in the educational literature (Hulleman & Cordray, 2009; Raudenbush & Sadoff, 2008; Smith, Daunic, & Taylor, 2007). In the past decade more emphasis has been placed on the importance of carefully measuring treatment implementation and using these measures in statistical analyses to improve the interpretation of study findings. With a heightened awareness of the need for improving measurement and use of treatment fidelity, researchers are grappling with the complexities of accomplishing this goal.

Components of Treatment Fidelity

Although the definition of treatment fidelity is straightforward, measuring it is complicated by the varied nature of treatments and logistical challenges faced in intervention research. Dane and Schneider (1998) identified five key aspects of treatment fidelity (i.e. integrity): (1) adherence, the degree to which the program components were delivered; (2) dosage, the amount of the intervention delivered, including number and length of sessions; (3) quality, general measures of quality of implementation, including implementer enthusiasm, preparedness, and positive attitudes; (4) participant responsiveness, the level of engagement including the degree of participation and enthusiasm; and (5) program differentiation, the extent to which a program can be distinguished from other similar programs. Durlak and Dupre (2008) drew attention to three additional aspects of fidelity: (6) control/comparison condition descriptions, (7) program reach (degree and scope of participation), and (8) adaptation, description of modifications made during implementation. Adequately addressing each of these components requires a theoretical understanding of the critical features of the intervention as well as of the comparison condition, which may not be specified or systematic. It may also require overcoming logistical obstacles, such as cost of technology (e.g. cameras for videotaping), or barriers imposed by schools (e.g. access to classrooms and information regarding instruction).

Past Trends in the Use of Fidelity Data

Fidelity measurement in educational intervention research has progressed substantially in recent years. Prior to the 1980's, few studies mentioned fidelity or related terms (i.e., treatment fidelity, program integrity, procedural reliability, or treatment checks; Gersten, Baker, & Lloyd, 2000; Smith, et al., 2007). In the late 1980's, educational researchers began to report fidelity in terms of procedures, training of intervention implementers, and other documentation that the intervention was implemented as intended (Dane & Schneider, 1998). Currently, researcher-created checklists or rubrics are typically employed to gauge whether or not aspects deemed critical to the intervention have occurred; however, fidelity data are often reported with little context regarding procedures or reliability of raters (Gersten, et al., 2000). A lack of uniform usage and debate about "acceptable" levels of fidelity reporting has become a salient issue (Smith, et al., 2007).

The practical use of fidelity data has been to monitor the implementation of the intervention while an experiment is ongoing. Ideally, those administering the treatment are well trained before the intervention begins and then fidelity measures demonstrate the success of that training (Barber, 1973). However, if the administration of the intervention is found lacking, the researcher may feel pressure to make quick decisions to salvage an experiment. She could take steps to correct the situation, such as by retraining of teachers, so that subjects can receive as much exposure the intervention as possible. The advantage of this would be to reduce the dilution of the treatment group with subjects who have not had a full "dose" of the treatment. Its disadvantage is that it produces a changing treatment, which is a threat to scientific validity and interpretation of the results.

Recent Trends in the Use of Fidelity Data

An alternative use for fidelity data that is now more common is to make it a part of the analysis of outcome data. Any subject who receives less than full exposure to the intervention can have his or her outcome data's contribution to the analysis discounted in some way. An extreme example of this approach is to separately analyze or even discard outcomes from subjects receiving the intervention with low fidelity, as in Hornbacher, Dretzke, Peterson, and Hickey (2008) and Hulleman and Corday (2009). An approach that preserves more information from the data is to treat the fidelity received by the subject as an explanatory variable, and use it as a covariate for reducing unexplained variance, as in Simmons et al. (2010), Munter (2010), and Vadasy and Saunders (2009). Both these approaches can help determine whether weak implementation is responsible for reduction in the impact of the innovation on outcomes.

The Institute of Education Sciences (IES) has specified categories of funding for its grants that focus attention on the development of interventions, testing their efficacy, and examining how they are brought to scale. In each of these types of research, treatment fidelity has a role. In development grants, funds are provided to create effective and reliable measures of the fidelity. Fidelity is also an important outcome variable, as a primary purpose of these studies is to determine whether implementing an intervention is feasible. In efficacy trials, the focus is on establishing and maintaining high degrees of fidelity in order to examine the success of the intervention in producing desired outcomes. In scale-up research, high fidelity remains a goal, yet variability of implementation is anticipated. Fidelity is an important outcome variable in these studies too, as they seek to determine if an intervention can be implemented as needed to produce desirable outcomes. In each phase of research, fidelity is a key to interpreting outcomes.

The Concept of Measurement Error and its Consequences

When fidelity is used in the analysis of treatment outcomes, either as a covariate to help explain outcomes or as a definition of treatment groups (high vs. low fidelity), it is implicitly assumed that fidelity is measured accurately. When it is not, a model that links outcomes to fidelity will be estimated inaccurately. It is well known from statistical theory that fitting a simple regression model in which the explanatory variable is afflicted with measurement error will yield an estimated regression coefficient that is attenuated, or biased toward zero (Fuller, 1987). As a result, the strength of the relationship between the explanatory and response variables will be underestimated on average. When the imperfectly measured explanatory variable is fidelity and the response is student outcome, the analyst can be misled into believing that the relationship between the two is weaker than it is, in the sense that a prediction of expected student outcome for a full implementation would be underestimated. This could cause one to underestimate the importance of the intervention effect.

What makes this observation relevant is that fidelity is hard to measure exactly. First, it is often assessed by observers. The observers' assessments may have low reliability, either because the measures are difficult to determine (e.g., a measure may require counting very frequent behaviors) or because they require subjective assessment. This lack of consistency in the measure, either within or among observers, will be referred to as "observer variance." Second, it is typical that not all intervention sessions are monitored for fidelity. If average fidelity over observed sessions is used as the measure, it will not necessarily be the same as the measure obtained if all sessions were observed. This results in uncertainty that will be referred to as "sampling variance". Together, these components comprise measurement error in the fidelity measure.

The motivation for this investigation came from findings in medical intervention studies, particularly in nutrition. Researchers noted that confirming a relationship between disease status and nutrient intake was more difficult in humans than in animals. They determined one reason to be that the measures of compliance to an intervention diet in humans (such as a low-fat diet) suffered from large measurement errors. As a result, the estimate of the strength of the relationship between the nutrient and health outcome was reduced (Kipnis, et. al. 1999).

Raudenbush and Sadoff (2008) discuss this phenomenon in an application to educational research. They studied the relationship between classroom quality and student outcome, where the quality measure suffered from low reliability. They noted that this caused the strength of the relationship between classroom quality and student outcomes to be underestimated. They described a method to remove the bias, taking into account both uncertainty in the quality measure and the hierarchical data structure.

This is similar to the scenario faced by researchers using fidelity as an explanatory variable in an analysis of outcome data. High quality studies acknowledge this fact by including assessments of their fidelity measures (e.g., estimated reliabilities) in their experimental protocol. However, as noted, this is not the only source of uncertainty in fidelity measures. Further, methods of accounting for this uncertainty in outcome analysis are not currently being used.

Purpose and Organization

The purpose of this paper is two-fold. First, we propose a simple method that uses fidelity to predict the maximum possible effectiveness an intervention might have if it had been fully implemented. This can benefit theory testing. Second, we show how this prediction should be modified to reduce bias when fidelity measurement is error-prone. The method of bias correction is the same as that in Raudenbush and Sadoff (2008) except that we illustrate a simpler implementation method known as regression calibration. We also extend their methods to consider the two sources of variability in the measurement of fidelity: observer and sampling variance.

The methods are illustrated by an application to data collected during a large-scale longitudinal intervention study, Project Maximize, whose findings are reported elsewhere

(Allor, Mathes, Roberts, Cheatham, & Al Otaiba, 2012; Allor, Mathes, Roberts, Cheatham, & Champlin, 2010; Allor, Mathes, Roberts, Jones, & Champlin, 2010). First, we describe Project Maximize and the process of collecting fidelity data in that study. Second, we discuss how measurement error arises in fidelity measures and how statistical techniques can be used to correct for it. Third, we present a summary of the fidelity data from Project Maximize. Fourth, the methods for integrating fidelity data into analyses and minimizing error are demonstrated through an analysis of Project Maximize data. Finally, we discuss how these techniques can be applied to development, efficacy, and scale-up research.

Intervention Study

Project Maximize Overview

The intervention study on which fidelity data was collected was a longitudinal randomized control trial examining the effectiveness of a comprehensive reading intervention for teaching students with IQs between approximately 40 and 79 to read (Allor, et al., 2012). Most intervention research in reading has been conducted with students with IQs of 80 or higher. This study explored how effectively methods proven successful with students having IQs near or above average ranges would also be successful with students having IQs below 80. The rate of progress was anticipated to be slower than average, so a longitudinal design was used; students were provided instruction for up to four years. All students were in grades 1-4 when they began the study; additional students joined the study in the second and third years. The final sample included 76 students in the treatment and 65 students in the contrast group. Students were excluded from the final sample if they participated less than one academic year. Students in the contrast group were provided with instruction according to the district and school expectations. Students in the treatment group participated in a comprehensive intervention that included all major components of reading instruction (i.e. phonemic awareness, listening comprehension, phonics, fluency, reading comprehension, and spelling). Instruction followed the techniques of Direct Instruction (Carnine, Silbert, Kame'enui, & Tarver, 2004; Coyne, Kame'enui, & Simmons, 2001; Engelmann, 1997), including systematic and explicit instruction in phonics. The primary program used was *Early Interventions in Reading,* also reported as *Proactive Reading* (Allor & Mathes, 2012; Mathes, et al., 2005; Mathes & Torgesen, 2005).

An extensive battery of reading measures was used, including annual standardized measures, as well as more frequent progress monitoring measures. Statistically significant differences between groups were found on all measures except one (untimed word recognition). The results indicated that (a) the comprehensive, structured reading intervention was effective in improving reading performance for students with IQs between 40 and 79; and (b) IQ had a statistically significant positive relationship with on student response to the intervention. Although the influence of IQ was a clear finding in the data, some students' performance did not fit this pattern. That is, some students with low IQs made faster gains than students with higher IQs. The most important implication of the study is that students with low IQs should be provided with systematic, explicit, comprehensive, and intensive reading instruction, techniques that are effective for other learners who struggle to learn to read (Allor, et al., 2012).

Fidelity in the longitudinal study was measured by observing each teacher two or three times each year using a fidelity instrument that assessed 8 teacher characteristics, each with a three-point rating scale (see Allor, et al., 2012). Although no formal measures of adherence to a curriculum were made in comparison classrooms, data about the type of reading curriculum implemented in these classrooms were collected and described, particularly about overlap with the instructional procedures of the treatment.

The Fidelity Study in Maximize

Fidelity was measured and the measures examined intensely during the final (fourth) year of the longitudinal study. The number of students participating in the study that year was 56 in the treatment group and 46 in the contrast group. Each of the 9 teachers who taught the intervention that year taught several groups (1-3 students per group) and was videotaped teaching each instructional group approximately five times. We randomly selected one videotape of each student in the treatment group to code for fidelity of implementation. If a video included more than one student, a second videotape was randomly selected for the additional student(s) and coded. Fidelity scores were linked to all students in each videotaped lesson; thus, students who were taught with other students had more than one fidelity score linked to their outcome data. The average number of coded sessions per teacher was $k = 9.5$. A sample of these tapes was recoded independently by an equivalently trained observer. On average, there were $r = 1.6$ double-coded sessions per teacher.

For this investigation, the original fidelity observation measures were refined. Initially, fidelity observations were conducted live without videotaping, but a sample of sessions was also videotaped. The videotapes allowed scrutiny of the fidelity instrument and improvements in its sensitivity and coding reliability. The original instrument included eight items that were rated from 0-3 on a Likert scale. These addressed several aspects of fidelity, including adherence, quality, and participant responsiveness. Specific items were appropriate pacing, adherence to lesson procedures, individual practice, error correction/scaffolding, student mastery of activities, student attentiveness, unison responses, and instructor warmth and enthusiasm. Initially, the same items were retained, but the Likert scale was expanded to 0-5 and more specific criteria were developed for determining the rating of each item. Each lesson included several activities, each of which was rated on these items.

After reviewing the descriptive data, 3 of the original items were omitted from this analysis. Two of these evaluated the teacher's use of individual practice and unison responses. Since the majority of groups in this final year of the project included only one or two students, the ratings on these items were uniformly high or coded as not applicable. The third item eliminated, instructor warmth and enthusiasm, exhibited almost no variation, as all the teachers received very high ratings on this item. Thus it contained no information about how fidelity on this item influenced treatment outcomes. This also was not a unique aspect of the treatment protocol, and no similar measure was available for the contrast group. The remaining items, referred to as aspects of fidelity, are described in Table 1, with a description of how the rating of each item was determined.

Although a battery of measures was used in the longitudinal study (see Allor, et al., 2012), for illustrating the methodology proposed in this paper, a measure of decoding fluency is used. The measure is the phonemic decoding efficiency subtest of the Test of Word Reading Efficiency (Wagner, Torgesen, & Rashotte, 1999), referred to here as "phonemic decoding." As with many measures of reading, the norming sample did not include students with intellectual disabilities; however the test has become a standard component of reading batteries and has strong reliability and validity.

Measurement Error in Fidelity

There are two potential sources of measurement error in fidelity measures, observer and sampling variance. Observer variance is often acknowledged in descriptions of fidelity measurement in educational experiments by a statement about its reliability. Observers are typically trained until they achieve an acceptable reliability. Research designs may also include monitoring of the consistency of observers during the intervention. Statistics such as reliability or agreement rates may be calculated from fidelity measurements made by two observers for at least a subset of the intervention sessions. These statistics are often reported, but rarely used other than to deem the measure adequate or inadequate. For example, Simmons et al. (2010) reports that "Reliability on double-coded tapes ranged from 0.75 to 1.0, representing an acceptable range for a moderate-inference instrument." (p. 134) Munter (2010) reports that he did not include in his analysis "any of the nine indicators for which coder agreement was less than 70%." (p. 70)

The second reason that a measure of fidelity may differ from its true value is sampling variance. Usually only a sample of the sessions in which the intervention is delivered are observed. The estimated mean fidelity for the sessions that are observed may differ from the true mean for all sessions if fidelity varies between sessions. In contrast to reliability, this variance and the uncertainty in estimation it causes is rarely acknowledged in educational experiments. One recent exception to this is Simmons et al. (2010). They signal the existence of variability in fidelity from one session to another by reporting a correlation between the measurements made on two different sessions of each teacher. (Actually the measure they were reporting on was teaching quality, which is different from but related to their index of fidelity.) They refer to this statistic as "stability of teaching quality," and report its value as between .60 and .70.

A useful way to describe uncertainty due to sampling variability in the fidelity measure is as a standard error. Specifically, the standard error of the average sample fidelity reflects its imprecision, as long as the sample of observed intervention sessions are (or can reasonably be thought of as) a random sample of all sessions. The standard error depends both on how much fidelity varies from session to session as well as the number of sessions in which fidelity is assessed.

The number of sessions observed for evaluating fidelity varies widely in studies. For example, Simmons, et al. (2010) measured fidelity for 2 out of 36 intervention sessions, while Hulleman and Cordray (2009) report a study in which every intervention session was evaluated for fidelity. In the latter case, no sampling variability remains, although observer variability might still be present.

Quantifying the Components of Measurement Error in Fidelity

In this section a method for estimating the size of these two components of measurement error is described. Suppose that a fidelity measurement is made for each teacher on at least two sessions of his or her intervention delivery. Further assume that for some sessions and teachers, replicate measurements are available from more than one observer. Denote by w_{ijr} the fidelity measure recorded for the j^{th} session of the i^{th} teacher (called the $(i,j)^{th}$ session) by observer r. Let k_i denote the number of teacher i's sessions, out of a total of K_i, on which fidelity measurements are available. If the primary observer, say observer p, made the measurement on all k_i observed sessions, then one estimator of fidelity for teacher *i* is the average over all the sessions observed by the primary observer for that teacher; i.e.,

$$\overline{w}_{ip} = \frac{1}{k_i}\sum_{j=1}^{k_i} w_{ijp}. \qquad (1)$$

In this case, only the primary observer's data are used for the fidelity measures, and replicate measures are used only for assessment of observer uncertainty. (Alternatively, all observers' data could contribute to the fidelity measures by using the average of the r_{ij} observers' measures for session (*i*,*j*) in the place of w_{ijp} in (1).)

To understand how to assess error in the fidelity measure, it is useful to explicitly describe and provide notation for the true fidelity that would be assessed if it were possible to do so. Let x_{ij} denote the true value of fidelity for session (*i*,*j*). Then teacher *i*'s true fidelity would be

$$\overline{X}_i = \frac{1}{K_i}\sum_{j=1}^{K_i} x_{ij}\ , \qquad (2)$$

which is the mean of the true fidelity over all sessions taught by the teacher, whether they were sampled for fidelity measurement or not.

To allow comparison of the two sources of variability of $\overline{w}_{ip}$ from $\overline{X}_i$, both sources must be reported on the same scale. A natural scale to choose is that of variance. The contribution to variance from sampling variability is straightforward, since the variance of a sample mean is well understood and easily estimated. But neither reliability nor agreement rate, common measures of observer uncertainty, are direct measures of variance. A link of either to variance can be made through the classical measurement error (CME) model (Carroll, Raymond, Ruppert, Stefanski, & Crainiceanu, 2006, p. 2) to describe observer variability.

Under the assumptions of the CME model, observer *r*'s reported fidelity is its true value plus a zero mean error; i.e.,

$$w_{ijr} = x_{ij} + e_{ijr}, \qquad (3)$$

where e_{ijr} is the error in observer *r*'s fidelity measure for session (*i*,*j*). The model assumes that the errors have constant variances and are independent of both the true value of fidelity x_{ij} and of one another. The errors are sometimes assumed to be normally distributed as well, leading to the normal model

$$e_{ij} \overset{iid}{\sim} \text{Normal}(0, \sigma_e^2)\,. \qquad (4)$$

This means that each observer's fidelity measurements are correct, on average, and equally precise.

For analysis of data from Project Maximize, we used estimator (1) and the CME model (3) – (4) for observer errors for the fidelity measurements. The appropriateness of this model will vary from one study to another, depending on the fidelity measures used and how they are collected. This model may require adaptation for some fidelity measurement systems. For example, the assumption that all observer errors in model (3) are independent of one another will not hold if more than one observer's data is included in the estimator in (1), since observations of the same observer will be correlated. In that case, the model and estimation procedure described below could be adapted to include an observer variance component. Another assumption of the CME model that may be violated in some studies is that of zero mean errors, since observers may make biased assessments. When fidelity is a judgment-based assessment, as in this case, the notion of "true" fidelity is elusive, and so will be *defined* as the mean assessment over observers, making this assumption true by definition. In cases in which the fidelity is fact based, such as the count or amount of time the teacher performs a certain behavior, this assumption may not hold. Finally, the assumption of normality of errors may not hold. In our application, the fidelity measures are based on Likert items, and so reported measures can't be exactly normally distributed. However, since our fidelity measures end up being averages of a large number of items, over both activities within a session and sessions, the central limit theorem makes the assumption reasonable.

Observer variance is defined as the increase in variance of teacher *i*'s fidelity measure due to the observers' errors. Sampling variance is defined as the variance that would be present in the fidelity measure if the observer were perfect, but not all sessions were sampled. Thus in our application, the observer variance is

$$\sigma_{oi}^2 = \sigma_{\overline{w}_{ip}|\overline{X}_i}^2 - \sigma_{\overline{x}_i|\vec{X}_i}^2 , \qquad (5)$$

where $\overline{x}_i = \frac{1}{k_i}\sum_{j=1}^{k_i} x_{ij}$ is the measure of fidelity that would have been reported for teacher *i* if $\sigma_e^2 = 0$ in (3); i.e., if there were no observer variance, but the same sample size k_i had been used. (The conditional notation is used because later we will consider a model that regards the teacher fidelity $\overline{X}_i$ as random.) The variance components for $\overline{w}_{ip}$ are (see Cochran (1977), p. 382) the observer variance

$$\sigma_{oi}^2 = E(\overline{w}_{ip} - \overline{x}_i)^2 = \frac{1}{k_i}\sigma_e^2 \qquad (6)$$

and the sampling variance

$$\sigma^2_{\bar{x}_i|\bar{X}_i} = E\left[(\bar{x}_i - \bar{X}_i)^2 \mid \bar{X}_i\right] = \left(1-\frac{k_i}{K_i}\right)\frac{1}{k_i}\sigma^2_{xi} \qquad (7)$$

where σ^2_{xi} denotes the variance of teacher *i*’s true fidelity from one session to the next. Thus the total variance is

$$\sigma^2_{\bar{w}_{ip}|\bar{X}_i} = \left(1-\frac{k_i}{K_i}\right)\frac{1}{k_i}\sigma^2_{xi} + \frac{1}{k_i}\sigma^2_e\,. \qquad (8)$$

If both σ^2_{xi} and σ^2_e were known, then the relative size of each component of $\bar{w}_{ip}$’s variance could be determined from (6) – (8). Since they are not, they must be estimated from the data. Methods for estimating these parameters are discussed in the next section.

Not all studies will have both sources of variance in their fidelity measures, nor do two measures in the same study necessarily have the same sources of variance. For example, Hulleman and Cordray (2009) report a study of motivation in which fidelity is measured at the student level, since each student controls his own adherence to the intervention. They defined two indices of fidelity, one measuring quantity (dosage), how many times the student participated, and one measuring quality, how completely they participated. Both of these indices were measured for every session in which each student participated. Thus there is no contribution to the fidelity variance due to sampling variance for either fidelity measure, since $k_i = K_i$ and thus from (7) its value is 0. There is also presumably no observer error for the dosage fidelity measure, since the count of participation sessions can be made accurately. However, the quality fidelity measure was based on observer judgment, and those judgment values were reported to have 81% and 88% agreement in their two described studies. Thus there is a contribution to the variance of the measured fidelity from the observer, although we can’t assess its variance due to the nonequivalence of agreement rate on a Likert scale and observer variance.

Estimating the Magnitude of the Two Components of Variance

In order to estimate σ_e^2 and σ_{xi}^2 in (8), it is necessary that two or more of each teacher's sessions ($k_i \geq 2$) are observed for fidelity and that a subset of the sessions have their fidelity measures replicated either by an equivalently precise independent or by a "perfect" observer. A perfect observer is known in the literature of measurement error as the gold standard, and may not actually be perfect, but just considerably better than the primary (typically cheaper) observer. In Maximize, videotapes of sessions were independently assessed for fidelity by two graduate student observers, each with equivalent training. Thus we assumed the two were equivalently precise replicates and present an estimation procedure for this scenario.

Suppose that session (i,j), $i = 1,\ldots,I$ and $j = 1,\ldots,k_i$ are coded for fidelity by r_{ij} equivalently precise observers; i.e., model (3) is assumed for all. Then an unbiased estimator of σ_e^2 is

$$\hat{\sigma}_e^2 = \frac{1}{r_{\cdot\cdot} - k_{\cdot}} \sum_{i=1}^{I} \sum_{j=1}^{k_i} \sum_{r=1}^{r_{ij}} (w_{ijr} - \overline{w}_{ij})^2 \qquad (9)$$

where $\overline{w}_{ij} = \sum_{r=1}^{r_{ij}} w_{ijr} \big/ r_{ij}$ and $r_{\cdot\cdot}$ and $k_{\cdot}$ indicate summation of the r_{ij}'s and k_i's over their indices. An unbiased estimator of σ_{xi}^2 is

$$\hat{\sigma}_{xi}^2 = \hat{\sigma}_{wi}^2 - \hat{\sigma}_e^2 \qquad (10)$$

where $\hat{\sigma}_{wi}^2$ is an estimator of the total variance of teacher i's error prone observations:

$$\hat{\sigma}_{wi}^2 = \frac{1}{k_i - 1} \sum_{j=1}^{k_i} (\overline{w}_{ijp} - \overline{w}_{ip})^2 \,. \qquad (11)$$

If the session-to-session variance of all teachers can be assumed constant, the deviations within teachers in (11) can be pooled to produce an estimator of the common variance. In Maximize, the teachers' session-to-session variability differed substantially, so that individual variances were estimated.

Summary of Project Maximize Fidelity Data

To illustrate how the two components of variance in a teacher's fidelity measure can be compared, and what might be learned by doing so, we examine data from Project Maximize. Five aspects of the intervention were assessed for fidelity by a primary observer several times for each teacher, as well as by a second observer for a subset of the sessions. Each teacher's fidelity measures were computed as shown in (1). The secondary observer's data was used for estimating the error variance as in (9), and the teacher's variance across sessions as in (10). The results are shown in Table 2. The upper panel displays $\overline{w}_{ip}$ for each aspect and teacher, where the nine teachers have been ordered from worst to best on the average of their 5 fidelity scores. The lower panel displays $\hat{\sigma}_{xi}$, which is a measure of the teacher's consistency in delivery of the intervention.

In Project Maximize, teachers were hired specifically to implement the intervention and received extensive support. Table 2 shows that fidelity was high, with an average over aspects and teachers of 4.1, where 5 was the maximum achievable score. A score of 4.0 indicates that the intervention was implemented as intended with only minimal alterations or omissions. Teachers did vary in fidelity, and some aspects of the intervention were more variable than others. The lowest fidelity was observed on the error correction item, indicating that errors were not corrected using the appropriate correction procedure, or they were not corrected at all. The highest scores were on attentiveness, indicating that students exhibited little off-task behavior.

Teachers varied more on consistency than average of fidelity. The teachers that were best were also most consistent. For example, Teacher C had the highest overall fidelity score and the smallest standard deviation across sessions on most items. Teachers H and I had the lowest overall scores and also the most variation across sessions. It is likely that this is partially an artifact of the Likert scale, which was bounded by 5; however, it could also be that quality and consistency are related.

The fidelity measures $\overline{w}_{ip}$ displayed in Table 2a are estimates of $\overline{X}_i$, with uncertainty due to both sampling and observer variance. Table 3 displays estimates of the total variance $\sigma^2_{\overline{w}_{ip}}$ for each teacher and fidelity measure. They are calculated by substituting the variance component estimates from (9) – (11) into the expression shown in (8). The percentages of the total variance explained by sampling variability (ratio of estimates of (7) to (8)) are also shown in Table 3.

The most notable observation from Table 3 is that nearly all of the uncertainty in the measure of fidelity in Project Maximize was due to sampling, rather than observer, variability. The reason for this is that most teachers' session to session variance was substantial (Table 2b), and the observers were very reliable. Only for the most consistent teacher (Teacher C) did observer variance make up a substantial fraction of uncertainty. The small fraction of total variance that observer variance makes up in this example is a reminder that researchers should think carefully about how to expend resources in their fidelity data collection. In this case, the uncertainty in all fidelity measures could have been reduced most effectively by increasing the size of the sample of sessions observed and not by trying to achieve greater observer consistency.

Using Fidelity in Estimation of Effect Size

In this section, we illustrate the benefit of viewing effect size as a function of fidelity. We also show the impact that imperfectly measured fidelity can have on estimation. Data from Project Maximize is used demonstrate these points. Three analyses are presented. The first ignores fidelity; the second takes it into account but ignores the variability in the fidelity measures, and the final analysis considers both the impact of fidelity and its measurement on estimation.

Model 1: Ignores fidelity

The analysis of intervention effects for most randomized trials in education allow different mean outcomes for treatment and contrast groups, along with a hierarchical structure, where students are nested within classrooms (or teachers). There may also be covariates available to explain some of the variability in outcomes not related to the intervention. In Project Maximize, randomization to the treatment or contrast group took place at the student level, and all students with a teacher were either in the treatment or contrast group. An important covariate of outcome was IQ. Thus the following model was fit:

$$y_{tis} = \beta_t + \beta_{tz} \cdot z_{tis} + \alpha_{ti} + \varepsilon_{tis}, \qquad (12)$$

where y_{tis} and z_{tis} are the outcome and IQ of the s^{th} student with the i^{th} teacher receiving treatment $t = T$ (treatment) or C (contrast); $i = 1,\ldots, n_t$, and, $s = 1,...,m_{ti}$ where n_t is the number of teachers in group t and m_{ti} is the number of students taught by teacher i. The teacher effects α_{ti} are considered random; assumptions of independence normality of teacher and student residuals were made; that is,

$$\alpha_{ti} \overset{iid}{\sim} N(0,\sigma^2_{t\alpha}) \text{ and } \varepsilon_{tis} \overset{iid}{\sim} N(0,\sigma^2_{t\varepsilon}), \qquad (13)$$

where α_{ti} and ε_{tis} are independent of each other. Under model (13), effect size is defined as the difference in the means of treatment and contrast scaled by the standard deviation of outcomes. We use the standard deviation of outcome in the contrast group in our estimator, as that of the treatment group should be biased upward by the varying fidelity of intervention. Thus we define effect size as

$$ES = \frac{(\beta_T - \beta_C) + (\beta_{Tz} - \beta_{Cz}) \cdot \bar{z}}{\sqrt{\sigma^2_{C\alpha} + \sigma^2_{C\varepsilon}}}. \qquad (14)$$

ES is estimated from the data by substituting estimates of parameters $\beta_T, \beta_C, \sigma^2_{C\alpha}, \sigma^2_{C\varepsilon}$ from the model in (12) and (13) into expression (14). This estimator is denoted by

Table 4 shows results of fitting the model in (12) to the gain in the phonemic decoding score for students enrolled during the last year of Project Maximize. An estimate of the effect size (as defined in (14)) is also shown. The analysis shows that virtually all (to two decimal places) of the variability in scores is at the student level for the contrast group, and that the effect size is 0.50, which could be described as medium.

Model 2: Treats fidelity as known but without error

One way to include fidelity data in the analysis is to use it an explanatory variable in the outcome model. Whether the fidelity measurement is individual to student (as in Hulleman and Cordray 2009) or teacher (as it more typically is) will determine the form the model will take. The contrast classrooms may not be evaluated for fidelity; indeed, the notion of fidelity to the contrast treatment is problematic since the components are likely not comparable to those of the intervention. In Maximize, fidelity was measured at the teacher level and only for the treatment group, so Model 2 reflects that.

Thus model (12) is expanded for outcomes of students in the treatment group to

$$y_{Tis} = \beta_T + \beta_{Tz} \cdot z_{Tis} + \beta_w \overline{w}_{ip} + \alpha_{Ti} + \varepsilon_{Tis}, \qquad (15)$$

where $\overline{w}_{ip}$ is a vector of average fidelities for teacher *i* and β_w is the vector of their regression coefficients. The outcomes for students in the contrast group are modeled as in (12) if there are no fidelity measures available for them, and analogously to (15) if there are. In either case, the assumptions of (13) still hold.

Under Model 2, effect size must be defined differently than previously, since the expected performance of the students in the treatment group will depend on the fidelity of their teachers. Depending on the purpose of the study, the goal may be theory testing, in which case the expected outcome of a perfectly implemented intervention is of interest, or scale-up, in which case the attained outcome, taking into account the failure of trained teachers to attain perfect implementation, may be the desired measurement. The model defined in (15) can be used for either notion of effect size. We describe the average gain of students for a perfectly implemented intervention as the maximum effect size,

$$MES = \frac{(\beta_T - \beta_C) + (\beta_{Tz} - \beta_{Cz}) \cdot \mu_Z + \beta_w w_{\max}}{\sqrt{\sigma^2_{C\alpha} + \sigma^2_{C\varepsilon}}}, \qquad (16)$$

where $w_{\max}$ is defined as the maximum possible (or maximum realistically possible) fidelity and μ_Z as the average of the covariate for students in the population. The attained effect size is defined as

$$AES = \frac{(\beta_T - \beta_C) + (\beta_{Tz} - \beta_{Cz}) \cdot \mu_Z + \beta_w \mu_W}{\sqrt{\sigma^2_{C\alpha} + \sigma^2_{C\varepsilon}}}, \qquad (17)$$

where μ_W is the average fidelity obtained for all the teachers. *MES* is larger than *AES* as long as fidelity and outcomes are positively related. *AES*, on the other hand, will be equivalent to *ES* under models (12) and (15). *MES* and *AES* can be estimated by substituting estimates from fitting Model 2 into expressions (16) and (17), μ_Z can be estimated by the average IQ of students in the study and μ_W by the average fidelity attained for teachers.

To illustrate interpretation of Model 2, we present results separately for measures of two aspects of fidelity from Project Maximize. This shows how the behavior of the model differs when the relationship between fidelity and outcome differs. The fidelity measure showing the largest impact on outcome was *Pacing*, while *Adherence* showed little. As in Model 1, IQ was included as a covariate. Table 5 shows estimates of the parameters of Model 2 (eqn. 15) for Project Maximize data, along with the effect sizes defined in (16) and (17), for the models using each fidelity measure. First note that comparing Tables 4 and 5 shows that adding fidelity as a covariate does not change the impact of IQ on reading outcome, since the regression coefficient of IQ is similar in the two models.

Observe that *MES* is substantially larger than *AES* for *Pacing*, but not for *Adherence*. The interpretation is that if all teachers were as faithful to the pacing required by the intervention as the best teacher, they would achieve an effect size that is conventionally regarded as large, while the same cannot be said for *Adherence*. The difference in the size of the estimated regression coefficients of the two fidelity measures is another way of quantifying the relative importance of the two aspects of fidelity. Increasing a teacher's *Pacing* measurement by one unit on the Likert scale would be expected to produce about five times as great an improvement in outcome as a similar improvement in *Adherence*. A caution in this interpretation is that the lack of a strong observed association between *Adherence* and outcome could be because all of them have achieved an adequate level, so that the association cannot be observed. Finally, observe that AES simply reflects the effect size at the level of fidelity that occurred in Project Maximize, so is the same as that in Model 1.

Model 3: Treats fidelity as known but imperfect

In this section we present the most general model considered for incorporating fidelity measurement into outcome analysis. Suppose that a fidelity measure is observed on a subset of k_i of the K_i intervention occasions. Furthermore, there may be variation among observers on fidelity assessment, which has been confirmed by replicate observations on either all or a subset of r_i of the k_i fidelity assessments for the i^{th} teacher. But now $\overline{w}_{ip}$ is regarded as only an estimator of the true fidelity $\overline{X}_i$, due to the presence of measurement error, which can arise from both sampling and observation errors. We assume the classical measurement error relationship between true and measured values of fidelity, as in (3).

The model we would like to fit to the treatment group is

$$y_{Tis} = \beta_T + \beta_{Tz} \cdot z_{Tis} + \beta_x \overline{X}_i + \alpha_{Ti} + \varepsilon_{Tis}. \tag{18}$$

However, the true fidelity $\overline{X}_i$ for teacher i is unobservable, so the usual method of estimation is not available. We use an alternative estimation procedure known as regression calibration, which is an algorithm that has been frequently applied in regression problems in which one or more of the predictors are observed with error. (See, for example, Carroll et al. (2006), Chapter 4.) The idea of the algorithm is that a linear predictor of $\overline{X}_i$ conditioned on the observed data, $E(\overline{X}_i \mid \overline{\mathbf{w}}, \mathbf{z})$,is first determined, where $\overline{\mathbf{w}}$ denotes the vector of teacher fidelity measures and $\mathbf{z}$ the vector of covariates that appear in the model. (In (18), IQ is the only covariate.) Then this predictor is estimated from the data and substituted for $\overline{X}_i$ in (18) before the model is fit as usual.

The appropriate predictor of $\overline{X}_i$ depends on assumptions about the distribution of fidelity and its measurement error. Under the CME model shown in (3) – (4), the best linear unbiased predictor (BLUP) for $\overline{X}_i$ is a weighted average of the fidelity measure for that teacher and the overall average fidelity adjusted for the covariates. (See Carroll et al (2006), p. 471, eqn. (4.4).) The BLUP contains unknown parameters that must be estimated from the data. These BLUP and its estimators are displayed in the Appendix in equations (A1) – (A6).

Regression calibration is a practical method for adjusting for measurement error in the predictors because once the BLUP for $\overline{X}_i$ is estimated and added to the file, model (18) can be estimated using any standard software package. The resulting estimators of the regression parameters in the hierarchical model are nearly unbiased and efficient (Buonaccorsi, Demidenko, & Tosteson 2000).

For Project maximize, we again illustrate the fitting of Model 3 for the two fidelity measures *Pacing* and *Adherence* separately. The results are shown in Table 6. Correcting for bias due to measurement error yields a slightly higher estimate of the coefficient of fidelity for *Pacing*, but not for *Adherence*. From Table 3, we see that the measurement error for *Adherence* is smaller than that for *Pacing*, resulting in less bias to be corrected. The adjustment of the coefficient for *Pacing* is upward in Model 3 because measurement error in a predictor attenuates the regression coefficient of the error prone variable, or biases it toward 0, when the predictor and other model covariates are uncorrelated. Since

IQ is only weakly correlated with fidelity, $\hat{\beta}_w$ from Model 2 is biased slightly downward for β_x.

Discussion

The purpose of this article was to (a) demonstrate a method for predicting maximum possible effect size (i.e. the effect size that theoretically could have been achieved if an intervention had been fully implemented) to account for variability in implementation; and (b) demonstrate methods for minimizing error variance by considering two sources of variability in the measurement of fidelity, observer and sampling variance. When these procedures are combined, researchers would be able to more accurately interpret their results as the variability common in fidelity data would be utilized in the calculation of the maximum effect size and error variance could be minimized by considering how fidelity data is collected. This article provides a theoretical rationale and demonstration of these procedures. The advantage for researchers is clear. Instead of collecting fidelity data for the purpose of demonstrating that an intervention was delivered with reasonably adequate fidelity, it enables the researcher to analyze how variation in fidelity impacts outcomes. This will not only lead to clearer understanding of whether the theory of the intervention (i.e. its active ingredients) improves outcomes, but also a more nuanced understanding of how the variation in aspects of fidelity changes outcomes. In turn, this will enable interventionists to modify and improve treatments and their implementation based on evidence that links fidelity data to outcomes. How these methods are applied within common intervention designs will vary according to the purpose of the research. The following paragraphs address how and why these techniques can be applied to the development of interventions, efficacy trials, and studies of intervention effectiveness when the intervention is applied on a large scale.

One important aspect of intervention development is the development of fidelity measures. In order to implement the techniques described in this article, measures of fidelity must be sensitive to variability in implementation and must represent the aspects of the intervention that the interventionist theorizes is causing improved outcomes (i.e. its active ingredients). Within the current goal structure of the IES grant application process, funds are provided to develop interventions, including the development of fidelity measures. This emphasizes the recognition within the field of the central role of fidelity in evaluating and understanding how and why interventions are effective (or ineffective). As we described in this article, increasing the sensitivity of fidelity measures is necessary to benefit from these techniques. The purpose of fidelity subtly changes from demonstrating that the intervention was implemented reasonably to measuring the variations in implementation. These variations will inform the iterative process of development. For example, in the early stages of developing an intervention, the measures of fidelity reveal the feasibility of the intervention, including which components are most challenging to implement. Further, the interventionist needs to take care that the measure is sensitive so that it can be fully utilized in both efficacy trials and wide scale application (i.e. scale-up research). Fortunately, with advances in technology and the ease of videotaping and improvements in electronic storage, it is becoming more realistic to refine fidelity measures as we did in our study.

With a sensitive measure of an intervention, a researcher conducting an efficacy trial will be able use fidelity data to calculate a maximum possible effect size and minimize observer and sampling variance as needed, depending on the context of the efficacy research. It is common in efficacy trials to have high degrees of fidelity to ensure that the impact of the intervention when implemented fully is determined. This was the case in the data described, as teachers were given intensive support and monitored carefully. However, with sensitive fidelity measures, even subtle differences in implementation may inform conclusions. In our example, when variation in pacing and adherence were incorporated into the maximum effect size, we could calculate the predicted effect of the intervention when these aspects were at their peak (i.e. the level obtained by our most effective teachers). For pacing, the effect size was 0.50, but the predicted effect size if implemented consistently with excellent pacing was higher than 0.80. This communicates clearly the importance of pacing in our intervention and the need to address pacing in professional development and monitoring of the intervention. On the other hand, incorporating data on adherence into the effect size increased the effect size only. The overall mean on this variable was a little higher than pacing and variance was smaller; therefore, calculating the maximum effect size did not seem to produce much more information. With increased variance, as in scale-up research, this calculation would likely be more informative.

The advantages of the techniques described are most evident in scale-up research when interventions are being implemented by practitioners in the field and when researchers often cannot control implementation as they do in efficacy trials. Indeed, this is the purpose of scale-up research as it is used to determine how effective an intervention is when it is applied on a broad scale in the field with minimal influence from the researcher. Without the techniques described low implementation is problematic; with these techniques, low implementation is actually informative, providing the researcher with evidence to demonstrate how variations in implementation impact outcomes.

In summary, these techniques provide great advantages to intervention researchers, but they also require fidelity measures that are highly sensitive. During the development of interventions, fidelity measures must also be developed that are driven by theory and precisely measure the aspects of the intervention that are believed to be the reason for changes in outcomes. With sensitive measures, improvements in the intervention can utilize fidelity data more systematically than is current typical practice. These measures can then be used during efficacy trials to ensure careful measurement of the critical components of an intervention. Finally, scale-up research becomes much more efficient and nuanced in that even data reflecting low fidelity could be utilized, increasing power and improving the interpretation of findings.

References

Allor, J. H., & Mathes, P. G. (2012). *Early interventions in reading: Level K*. Columbus, OH: SRA/McGraw-Hill.

Allor, J. H., Mathes, P., Roberts, K., Cheatham, J. P., & Al Otaiba, S. (2012). Is scientifically-based reading instruction effective for students with Below-Average IQs? Manuscript submitted for publication.

Allor, J. H., Mathes, P. G., Roberts, J. K., Cheatham, J. P, & Champlin, T. M. (2010). Comprehensive reading instruction for students with intellectual disabilities: Findings from the first three years of a longitudinal study. *Psychology in the Schools, 47*, 445-466.

Allor, J. H., Mathes, P. G., Roberts, J. K., Jones, F. G., & Champlin, T. M. (2010). Teaching students with moderate intellectual disabilities to read: An experimental examination of a comprehensive reading intervention. *Education and Training in Autism and Developmental Disabilities, 45,* 3-22.

Barber, T. (1973). Pitfalls in research: Nine investigator and experimenter effects. In R.M.W. Travers (Ed.), *Second handbook of research on teaching* (pp. 382-404). Chicago: Rand McNally.

Buonaccorsi, J., Demidenko, E., and Tosteson, T. (2000). Estimation in Longitudinal Random Effects Models with Measurement Error. *Statistica Sinica, 10*, 885-903.

Carnine, D. W., Silbert, J., Kame'enui, E. J., & Tarver, S. G. (2004). *Direct instruction reading* (4th ed.). Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

Carroll, Raymond J., D. Ruppert, L. Stefanski, C. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, 2nd Edition*, Chapman and

Hall.

Cochran, W.G. (1977) *Sampling Techniques*, 3rd edition John Wiley and Sons.

Coyne, M. D., Kame'enui, E. J., & Simmons, D. C. (2001). Prevention and intervention in beginning reading: Two complex systems. *Learning Disabilities Research and Practice, 16*(2), 62-73.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*(1). 23-45.

Engelmann, S. (1997). *Preventing failure in the primary grades*. Eugene, OR: Association for Direct Instruction.

Fuller, Wayne (1987) Measurement Error Models, New York: John Wiley and Sons.

Gall, Meredith D., Gall, J. P., & Borg, W. R. (2007). *Educational research.* Boston: Pearson/Allyn & Bacon.

Gersten, R., Baker, S., & Lloyd, J. W. (2000). Designing high-quality research in special education: Group experimental design. *The Journal of Special Education*, *34*(1), 2-18.

Hornbacher, M.J., Dretzke, B.J., Peterson, K.A., and hickey, M.C. (March, 2008). Looking more deeply: Fidelity of implementation as a critical component in evaluating intervention impacts. Paper presented at the Annual Meeting of the American Educational Research Association.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*(1), 88-110.

Kipnis, V., Carroll, R. J., Freedman, L. S., & Li, L. (1999). Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies, *American Journal of Epidemiology*, *150*(6), 642-651.

Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly, 40*, 148-182.

Mathes, P. G., & Torgesen, J. K. (2005). *Early interventions in reading, Level 1*. Columbus, OH: SRA/McGraw-Hill.

Munter, C. (2010, August). *Evaluating math recovery: The impact of implementation fidelity on student outcomes* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (AAT 3430739).

Raudenbush, S.W. and Sadoff (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness, 1*(2), 138 – 154.

Simmons, D., Hairrell, A., Edmonds, M., Vaughn, S., Larsen, R., Willson, V.,… Byrns, G. (2010). A comparison of multiple-strategy methods: Effects on fourth-grade students' general and content-specific reading comprehension and vocabulary development. *Journal of Research on Educational Effectiveness*, *3*(2), 121-156. doi:10.1080/19345741003596890

Vadasy, P. F., & Saunders, E. A. (2009). Supplemental fluency intervention and determinants of reading outcomes. . *Scientific Studies of Reading, 13*(5), 383-425. doi:10.1080/10888430903162894

Wagner, R., Torgesen, J., & Rashotte, C. (1999). *Comprehensive Test of Phonological Processing*. Austin, TX: PRO-ED.

Table 1

Aspects of Fidelity

Aspect	Explanation
1. pacing	materials were ready, pacing throughout the lesson was fast enough to keep students attention without being so fast as to increase errors, and teachers moved from one item/activity quickly
2. adherence	materials, techniques, and objectives were adhered to throughout the lesson
3. error correction/ scaffolding	all student errors were corrected according to prescribed correction procedure
4. mastery	students provided correct responses during initial presentation or when given an opportunity to correct an answer (i.e. responded correctly during correction procedure)
5. attentiveness	students exhibited on-task behavior throughout the lesson

Table 2

Fidelity and its Consistency for the Teachers of Project Maximize

a. Estimates of mean fidelity ($\bar{\mathbf{w}}_{ip}$) for each teacher

Teacher	Pacing	Adherence	Error Correction	Mastery	Attentive-ness	Average
I	2.96	3.79	3.14	2.84	3.55	3.49
B	3.51	4.00	2.94	3.46	4.03	3.75
A	3.73	3.66	3.66	3.70	4.68	3.98
E	4.07	3.58	3.57	4.17	4.81	4.15
D	4.36	3.84	3.39	3.99	4.69	4.20
F	4.19	4.06	3.82	4.12	4.31	4.20
H	3.45	4.25	3.58	4.75	4.70	4.29
G	3.95	4.68	3.68	4.65	4.67	4.44
C	4.28	4.61	4.42	4.71	4.95	4.65
Average	3.83	4.05	3.58	4.04	4.49	4.13

b. Estimates of standard deviation of fidelity ($\hat{\sigma}_{xi}$)for each teacher

Teacher	Pacing	Adherence	Error Correction	Mastery	Attentive-ness	Average
I	0.60	0.61	1.00	0.59	1.27	0.72
B	0.42	0.36	0.40	2.11	1.05	0.82
A	0.50	0.23	0.46	0.39	0.14	0.38
E	0.48	0.87	0.24	0.63	0.05	0.44
D	0.21	0.28	1.25	0.55	0.16	0.41
F	0.15	0.16	0.30	0.38	0.33	0.25
H	1.26	0.29	0.27	0.17	0.22	0.37
G	0.20	0.07	0.64	0.05	0.19	0.19
C	0.24	0.13	0.19	0.03	0.01	0.11
Average	0.45	0.33	0.22	0.53	0.38	0.41

Table 3

Sources of Measurement Error in Fidelity Measures

Teacher	Pacing		Adherence		Error correction		Mastery		Attentiveness	
	tot var	% sampling var.	tot var	% sampling var.	tot var	% sampling var.	tot var	% sampling var.	tot var	% sampling var.
I	0.66	91%	0.65	94%	1.04	96%	0.67	88%	1.28	99%
B	0.48	87%	0.40	90%	0.45	90%	2.19	96%	1.06	99%
A	0.56	89%	0.27	86%	0.56	91%	0.47	83%	0.15	94%
E	0.54	89%	0.91	96%	0.28	83%	0.71	89%	0.06	84%
D	0.27	78%	0.32	88%	1.29	97%	0.63	88%	0.17	94%
F	0.21	71%	0.20	81%	0.35	87%	0.46	83%	0.34	97%
H	1.32	95%	0.33	88%	0.31	85%	0.25	69%	0.23	96%
G	0.26	77%	0.11	66%	0.69	93%	0.13	39%	0.20	95%
C	0.30	80%	0.17	77%	0.23	80%	0.11	28%	0.02	38%
Avg	0.51	84%	0.37	85%	0.58	89%	0.62	74%	0.39	88%

Table 4

Model 1 Parameter Estimates for Phonemic Decoding Gain Score

$\hat{\beta}_T$	$\hat{\beta}_C$	$\hat{\beta}_{Tz}$	$\hat{\beta}_{Cz}$	$\sigma^2_{C\alpha}$	$\sigma^2_{C\varepsilon}$	$\widehat{ES}$
4.70	2.31	0.17	0.11	0.00	22.85	0.50

Table 5

Model 2 *Estimates for Phonemic Decoding Gain Score*

Fidelity measure	$\hat{\beta}_T$	$\hat{\beta}_{Tz}$	$\hat{\beta}_w$	$\widehat{MES}$	$\widehat{AES}$
Pacing	-17.32	0.16	3.10	0.83	0.50
Adherence	-8.83	0.18	0.58	0.54	0.50

Table 6

Model 3 Estimates for Phonemic Decoding Gain Score

Fidelity measure	$\hat{\beta}_T$	$\hat{\beta}_{Tz}$	$\hat{\beta}_x$	$\widehat{MES}$	$\widehat{AES}$
Pacing	-18.36	0.15	3.44	0.85	0.48
Adherence	-8.69	0.18	0.55	0.58	0.48

Appendix

The best linear unbiased predictor of $\overline{X}_i$ given the vector of measured fidelities and the IQs for students is

$$E(\overline{X}_i \mid \overline{\mathbf{w}}, \mathbf{z}) = \mu_{\overline{X}} + \begin{pmatrix} \sigma^2_{\overline{X}} & \sigma_{\overline{X},z} \end{pmatrix} \begin{pmatrix} \sigma^2_{\overline{X}} + \sigma^2_{\overline{w}_{ip} \mid \overline{X}_i} & \sigma_{\overline{X},z} \\ \sigma_{\overline{X},z} & \sigma^2_z \end{pmatrix}^{-1} \begin{pmatrix} \overline{w}_{ip} - \mu_{\overline{W}} \\ \overline{z}_i - \mu_z \end{pmatrix} \qquad \text{(A1)}$$

where $\sigma^2_{\overline{w}_{ip} \mid \overline{X}_i}$ is defined in (8) and $\mu_{\overline{X}} = \mu_{\overline{W}}$ is mean fidelity over all teachers, $\sigma^2_{\overline{X}_i}$ its variance, and $\sigma_{\overline{X},z}$ the covariance between a student's IQ and teacher fidelity. The parameters μ_z and $\hat{\sigma}^2_{z_i}$ denote the mean and variance of IQ's of the population of students. In order to calculate a prediction from this expression, the means, variances and covariances must be estimated. The estimator $\hat{\sigma}^2_{\overline{w}_{ip} \mid \overline{X}_i}$ is provided in (8) – (11). The remaining parameters can be estimated by:

$$\hat{\mu}_{\overline{X}} = \hat{\mu}_{\overline{W}} = \sum\nolimits_{i=1}^{k} m_{Ti} \overline{w}_{ip} / m_T \ , \qquad \text{(A2)}$$

where $m_T = \sum\nolimits_{i=1}^{k} m_{Ti}$ is the total number of students in the experiment;

$$\hat{\mu}_z = \sum\nolimits_i \sum\nolimits_s z_{Tis} / m_T \ ; \qquad \text{(A3)}$$

$$\hat{\sigma}^2_z = \frac{1}{m_T - 1} \sum\nolimits_i \sum\nolimits_s (z_{Tis} - \hat{\mu}_z)^2 \ ; \qquad \text{(A4)}$$

$$\hat{\sigma}^2_{\overline{X}} = \hat{\sigma}^2_{\overline{W}} - \left(m_T \sum\nolimits_i m_{Ti} \hat{\sigma}^2_{\overline{w}_{ip} \mid \overline{X}_i} - \sum\nolimits_i m^2_{Ti} \hat{\sigma}^2_{\overline{w}_{ip} \mid \overline{X}_i} \right) \Big/ \upsilon_T \qquad \text{(A5)}$$

where $\hat{\sigma}^2_{\overline{w}} = \dfrac{m_T}{\upsilon_T} \sum\nolimits_{i=1}^{I} m_{Ti} \left(\overline{w}_{ip} - \hat{\mu}_{\overline{W}} \right)^2 / m_T$ and $\upsilon_T = m^2_T - \sum\nolimits_i m^2_{Ti}$;

and

$$\hat{\sigma}^2_{\bar{X},z} = m_T \sum_i \sum_s \left(\bar{w}_{ip} - \hat{\mu}_T\right)\left(z_{Tis} - \bar{z}_T\right) \Big/ \nu_T \,. \quad \text{(A6)}$$