Psychological Methods

A Power Analysis for Fidelity Measurement Sample Size Determination

--Manuscript Draft--

Manuscript Number:	
Full Title:	A Power Analysis for Fidelity Measurement Sample Size Determination
Article Type:	Article (unmasked review requested)
Abstract:	The importance and complexity of assessing fidelity has been emphasized in recent years, with increasingly sophisticated definitions of fidelity (Dane & Schneider, 1998; Durlak & Dupre, 2008; O'Donnell, 2008) and recommended procedures for developing fidelity instruments and collecting fidelity data (Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012). Researchers agree that in order to better understand an intent-to-treat study, measurement should be spread across the entire study period (Gersten, Baker, & Lloyd, 2000; Nelson, et al. 2012); however, little guidance has been provided about how to determine the number of observations needed to precisely measure fidelity (Smith, Daunic, & Taylor, 2007). With limited resources for research, this is an important question, particularly for interventions that last a considerable length of time. Increasingly, these data are being used to enhance the analysis of outcomes. This paper proposes a method for determining a reasonable sample size for fidelity data collection, in the case that fidelity assessment requires observation and coding of instructional sessions either live or by videotape. The proposed methodology is based on consideration of the power of tests of the treatment effect of outcome itself, as well as fidelity's contribution to the variability of outcomes. Software for the sample size calculation is provided.
Keywords:	power, measurement error, sampling variance
Corresponding Author:	Sara Lynne Stokes, Ph.D. Southern Methodist University Dallas, TX UNITED STATES
Corresponding Author E-Mail:	slstokes@smu.edu
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Southern Methodist University
Other Authors:	Jill Allor, Ed.E.
Corresponding Author's Secondary Institution:	
First Author:	Sara Lynne Stokes, Ph.D.
Order of Authors Secondary Information:	
Manuscript Region of Origin:	UNITED STATES
Suggested Reviewers:	Jessaca Spybrook Western Michigan University jessaca.spybrook@wmich.edu Our paper discusses how to use her software, OD, for sample size calculations
Opposed Reviewers:	

December 12, 2014

Professor Lisa Harlow
University of Rhode Island
Kingston, Rhode Island 02881

Dear Professor Harlow:

I am submitting, on behalf of myself and my co-author Professor Jill Allor, the attached manuscript for consideration for publication in *Psychological Methods*. We request that the paper receive an unblinded review. The paper addresses the question of how to select the sample size for fidelity measurement.

The manuscript has not been submitted for review to any other journal. It includes no primary data. Proper ethical guidelines have been followed in the conduct of the research and preparation of the manuscript.

We have prepared SAS computer code for implementing the procedure proposed. It has been prepared and submitted in a separate file from the manuscript. We request that it be considered as additional content.

Sincerely,

Lynne Stokes
Professor, Department of Statistical Science
Southern Methodist University

A Power Analysis for Fidelity Measurement Sample Size Determination

Lynne Stokes and Jill H. Allor

Southern Methodist University

Author Contact Information:

Corresponding author:
Lynne Stokes
Department of Statistical Science
Southern Methodist University
P.O. Box 750332
Dallas, TX 75275-0332
214-768-2270
214-768-4035 (fax)
slstokes@smu.edu

Jill H. Allor
Department of Teaching and Learning
Simmons School of Education and Human Development
Southern Methodist University
P.O. Box 750381
Dallas, TX 75275-0381
214-768-4435
214-768-2171 (fax)
jallor@smu.edu

This research was supported by a grant (R305D090020) from the Institute of Education Sciences.

Abstract

The importance and complexity of assessing fidelity has been emphasized in recent years, with increasingly sophisticated definitions of fidelity (Dane & Schneider, 1998; Durlak & Dupre, 2008; O'Donnell, 2008) and recommended procedures for developing fidelity instruments and collecting fidelity data (Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012). Researchers agree that in order to better understand an intent-to-treat study, measurement should be spread across the entire study period (Gersten, Baker, & Lloyd, 2000; Nelson, et al. 2012); however, little guidance has been provided about how to determine the number of observations needed to precisely measure fidelity (Smith, Daunic, & Taylor, 2007). With limited resources for research, this is an important question, particularly for interventions that last a considerable length of time. Increasingly, these data are being used to enhance the analysis of outcomes. This paper proposes a method for determining a reasonable sample size for fidelity data collection, in the case that fidelity assessment requires observation and coding of instructional sessions either live or by videotape. The proposed methodology is based on consideration of the power of tests of the treatment effect of outcome itself, as well as fidelity's contribution to the variability of outcomes. Software for the sample size calculation is provided.

Keywords: fidelity, power, measurement error, sampling variance

A Power Analysis for Fidelity Measurement Sample Size Determination

Fidelity of implementation means the extent to which theoretically meaningful components of an intervention are implemented as intended (Gall, Gall, & Borg, 2007). If a randomized trial of a promising intervention shows that an effect is small, or even non-significant, one possible explanation is that its important components were not fully implemented. The importance and complexity of assessing fidelity has been emphasized in recent years, with increasingly sophisticated definitions of fidelity (Dane & Schneider, 1998; Durlak & Dupre, 2008; O'Donnell, 2008) and recommended procedures for developing fidelity instruments and collecting fidelity data (Munter, Wilhelm, Cobb, & Cordray, 2014; Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012). Researchers are urged to be careful and complete when measuring both causes (i.e. independent variables) and outcomes (i.e. dependent variables) in their studies (Nelson, et al., 2012). In 1998, Dane and Schneider (1998) identified five aspects of fidelity: adherence, or following intervention protocol; exposure, or dosage; quality of delivery; participant responsiveness; and program differentiation, or how much the treatment condition varied from the comparison condition. Researchers are now being urged to more systematically and thoroughly measure interventions in order to (a) ensure that they are measuring the underlying theory that hypothesizes why an intervention may be effective and (b) more precisely measure variation in implementation (Nelson et al., 2012). Further, the concept of fidelity is even being expanded to include measuring variables that determine implementation (Nelson et al., 2012). In addition to conceptualizing fidelity more fully and operationalizing those definitions in measures, researchers are now using variability of implementation to better understand treatment effects by including them as covariates in their analyses. If sufficiently precise measures of fidelity of those components are available, the researcher may be able to determine whether non-significance is due to failure of the theory or inadequate implementation, and if outcomes improve with fuller implementation. With precise data, researchers may also be able to determine how to improve implementation in the future or to improve an intervention itself.

More sophisticated notions of fidelity present researchers with practical challenges. Each aspect of fidelity of implementation can be measured in a variety of ways. Some aspects of fidelity can be measured precisely. For example, suppose the measure is one of “dosage,” such as the number of times a student was present for a tutoring session, or the number of activities he or she completed. Then the fidelity measure may not require sampling or subjective assessment, and thus can be considered to be precise, given accurate recordkeeping. But dosage might also be thought of as opportunities to respond or the amount of practice of specific skills within a given time frame. One intervention, for example, might increase the dose by increasing the number of opportunities through faster teacher pacing or group responses rather than simply extending the length or frequency of sessions. Determining the number of opportunities to respond would require observation, which is more expensive than basic recordkeeping. With finite resources, researchers must be able to determine the number of observations required to adequately estimate fidelity. Researchers agree that in order to better understand an intent-to-treat study, measurement should be spread across the entire study period (Gersten, Baker, & Lloyd, 2000; Gersten, Fuchs, Compton, Coyne, Greenwood, & Innocenti, 2005; Nelson, et al. 2012); however, very little guidance has been provided about how to determine the number of observations needed to precisely measure fidelity (Smith, et al., 2007). With limited resources for research, this is an important question, particularly for interventions that last a considerable length of time. The resources required for fidelity have also increased with advances in video capability and the sophistication of observation instruments that often require multiple viewings of intervention sessions to complete reliably. Further, as interventions are scaled-up, large numbers of participants are included in studies, further increasing the number of potential observations that must be carefully coded and analyzed. Measuring fidelity of implementation in intervention studies can be just as costly as measuring outcomes (Gersten, et al., 2000). In this paper we address this challenge by proposing a method to determine the number of fidelity measurements needed for the goal of using it as a covariate in the analysis of outcome. The approach taken is to relate the number of measurements to the power of tests of treatment outcomes.

When designing randomized control trials, researchers are accustomed to making sample size decisions based on consideration of power. It has become routine to conduct power analyses to determine the number of subjects and clusters (e.g., classrooms, hospitals, or training centers) needed to achieve the desired power to test an intervention. Similarly, when fidelity measures are used in the analysis of outcome data, their precision also affects the power of the treatment effect, as well as the effect of the relationship of fidelity to outcome, which we will call the implementation effect. The power of both tests decreases as the variance of the fidelity measure increases. The purpose of this paper is to describe a principled approach, based on power, to determine the sample size needed to estimate fidelity, particularly in large scale-up studies.

In the next section, the relationship between the sample size and variance of the fidelity measure is discussed. Then in the following sections, expressions for the power of tests of the treatment and implementation effects are displayed for two designs: a person randomized and a two-stage cluster randomized design. These expressions show how to link power to the fidelity measurement sample size. Examples of two intervention studies from educational research show how to use the methods proposed for planning fidelity data collection. Finally, the last section provides a discussion of the implications for practice. Power of tests of treatment and implementation effects depends on a variety of characteristics of the study. Some of these will be known to the researcher, but others may be unfamiliar. We make recommendations about what data should be collected and reported so that adequate and efficient sample sizes can be more accurately predicted for future studies. Software (SAS code) for implementing the methods discussed is provided in supplementary materials.

The Variance of Fidelity Measures

The cases considered here are those in which the intervention is delivered in a series of discrete sessions, such as a tutoring session in an educational intervention or other kinds of training sessions. We suppose the fidelity measure must be assessed by observation of intervention sessions, either contemporaneously or by recording and subsequently evaluating the sessions. When only a subset of the sessions are assessed, the resulting estimate of fidelity may differ from the true fidelity that would have been computed had all sessions been observed.

For example, suppose the fidelity measure includes an item for the number of times a student or teacher exhibits some specific behavior in the classroom during the intervention period. Fidelity may be operationalized as the average number of times the behavior is observed per monitored session, or scaled up as this average times the number of times the student was present. But neither measure would be perfectly correlated with the true number of times the behavior occurred over the course of the intervention, unless the count of the behavior did not vary at all over sessions or all sessions were observed. However, if the sessions that are observed were selected randomly, or could be treated as such, then the estimator of fidelity would be unbiased for true fidelity and the size of its error will be reflected by its standard error.

To make that concrete, consider a person-randomized trial in which each person participates in *K* sessions over the course of the intervention study. Let f_{ij} denote the fidelity measure for person *i* in session *j* of the *K* sessions. The true fidelity for person *i* is defined as the average fidelity he or she received over all sessions:

$$f_i = \frac{1}{K}\sum_{j=1}^{K} f_{ij} \ . \tag{1}$$

(Fidelity could be defined as the total rather than the average if so desired, with appropriate changes in the expressions. The power of the designs remains the same for either definition.) If the fidelity measure were observed in only *k* of the *K* sessions, f_i could be estimated by

$$\hat{f}_i = \frac{1}{k}\sum_{j=1}^{k} f_{ij} . \tag{2}$$

If the sessions have been sampled randomly, $\hat{f}_i$ is an unbiased estimator of f_i and its variance is

$$V(\hat{f}_i \mid f_i) = \left(1 - \frac{k}{K}\right)\frac{1}{k}\sigma_{wi}^2 , \tag{3}$$

where σ_{wi}^2 is the within-person (session-to-session) variance in fidelity for person *i* (see, e.g., Lohr (2010), p. 36).

Thus the measure of fidelity (2) is related to true fidelity (1) by

$$\hat{f}_i = f_i + u_i , \tag{4}$$

where u_i has mean 0 and variance shown in (3), with f_i and u_i independent. When f_i and u_i (and therefore $\hat{f}_i$) are assumed to be normally distributed, model (4) is sometimes called the classical measurement error (CME) model (Carroll, Ruppert, Stefanski & Crainiceanu 2006, p. 2). The reliability of a measure that follows the CME model is the proportion of its between-person variance that is “real” variance; i.e.,

$$\kappa = \frac{Var(f_i)}{Var(\hat{f}_i)} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_u^2} \ . \qquad (5)$$

In the case of fidelity estimation, it is reasonable to assume $\hat{f}_i$ to be normally distributed (because it is an average). However, f_i will not necessarily be normal, depending on the fidelity measure chosen. Nevertheless, for purposes of fidelity sample size planning, we assume the CME model to hold, and that the within-person variance σ_{wi}^2 is constant (= σ_w^2) for all persons.

With these assumptions, the reliability for the fidelity measure when estimated from a sample of *k* sessions is (from (3) and (5)):

$$\kappa_k = \frac{\sigma_f^2}{\sigma_f^2 + (\sigma_w^2 / k)(1 - k/K)} = \frac{k * ICC_f}{k * ICC_f + (1 - ICC_f)(1 - k/K)}, \qquad (6)$$

where $ICC_f = \sigma_f^2 / (\sigma_f^2 + \sigma_w^2)$ is the fraction of the total variance of fidelity ($\sigma_f^2 + \sigma_w^2$) that is between persons.

ICC_f is denoted as such because it is an intra-cluster correlation, where the cluster is the unit (student or teacher) on which fidelity is measured. ICC_f measures consistency in the fidelity received from session to session. If fidelity is perfectly consistent over sessions, then the variance between sessions σ_w^2 = 0 and ICC_f = 1. This makes it much easier (and cheaper) to measure fidelity, since one measurement per individual would accurately reveal the true value of fidelity over the entire intervention period. Then the reliability of the fidelity measure would be perfect (κ_k = 1), as demonstrated by setting ICC_f = 1 in (6). On the other hand, if the individuals are identical to each other on average with respect to fidelity, but they are inconsistent from session to session ($\sigma_w^2 > 0$, $\sigma_f^2 = 0$, $ICC_f = 0$), then any attempt to distinguish fidelity between individuals by sampling is pointless because any observed differences are due simply to sampling variability, and this is reflected in the reliability ($\kappa_k = 0$). Note also from (6) that if every session were measured for fidelity ($k = K$), then reliability is also perfect, since each individual’s fidelity will be known exactly.

We can think of the sample size problem as determining how much reliability one should “buy.” It is not a one-size-fits-all decision, since it will depend on the consistency among and within individuals receiving the intervention. It will also depend on how the analyst intends to use the fidelity measures. As noted in the introduction, for [certain] types of studies, fidelity measures are used in the analysis of outcomes, commonly by including them as covariates in a regression model. There are two potential benefits from this type of analysis. First, if fidelity is related to outcome, it can improve power of a test of the treatment effect by reducing the unexplained variance in outcome. Second, detection of a positive relationship between fidelity and outcome provides some evidence of the correctness of the theory underlying the intervention. If implementation varies in an unplanned way, the argument for causation is stronger if better outcomes are associated with fuller implementation; i.e., if the data confirm the dose-response criteria of Hill (1953). Estimates of the fidelity slopes in different models can also provide information about the relative importance to different outcomes, or about the importance of different fidelity measures, or different items on an observation measure.

The reason that reliability of the fidelity measure is relevant for the study of power of the treatment effect in a randomized trial is that measurement error in any predictor of a regression model affects inference for all the parameters of the model. So if the fidelity measure is to be used as a covariate in the analysis of outcomes, its properties must be considered. The effects of measurement error in the independent variables of a regression model have been extensively examined, especially for the CME model. A summary of its effects follows (Fuller 1987):

(1) The estimator of the slope coefficient for the error-prone covariate is biased;

(2) The estimators of coefficients of other predictors will not be biased for predictors that are independent of the error prone one; but

(3) The residual variance of the model is increased, which reduces the power to detect non-zero coefficients for all predictor variables in the model.

The size of the bias, the increase in residual variance, and the reduction in power are all functions of the reliability of the error prone variable, as defined in (6). In the subsequent sections, we show how to use the relationship between power of the hypothesis tests of the model parameters and the sample size *k* to determine an adequate design for the collection of fidelity data. We will consider both person-randomized and group-randomized designs, a dichotomy also used by Spybrook (2012) for discussion of power in the OD software.

Relationship Between Fidelity Measurement and Power for Person-Randomized Single-Level Trials

The first study design considered is one in which *N* persons are assigned at random and individually (i.e., not as part of a class) to one of two treatments, where typically one is a control. These subjects receive the treatments in *K* discrete sessions. Fidelity is assumed to be measured in *k* of those sessions for each person in the experimental group. Current recommendations for systematic assessment of fidelity include determining to what degree the treatment condition varies from the control condition. This will typically include observation in the control group, sometimes using the same or a similar observation instrument. The question we address is what *k* should be.

Analysis of Outcomes Without Use of Fidelity

To set the stage for examining the value of high reliability in fidelity measurement, we first review a typical analysis of treatment effect that does not make use of the measure. To simplify expressions, we assume all designs are balanced, meaning that the same number of individuals is assigned to each treatment group. A typical model for outcome for person *i*, denoted by Y_i, in a person-level randomized trial is

$$Y_i = \beta_0 + \beta_1 W_i + r_i \,, \tag{7}$$

where the residual $r_i \sim N(0,\sigma^2)$ for $i = 1,\ldots, N$ and the number of subjects in each treatment group is $N/2$. W_i is defined as ½ or –½ according to whether subject *i* is in the experimental or control group, respectively. To test whether or not there was an effect from the intervention, the hypotheses would be set up as

$$H_0\text{: } \beta_1 = 0 \text{ against } H_a\text{: } \beta_1 \neq 0. \tag{8}$$

The *F*-test for testing (8) is based on the ratio of the mean squares:

$$F_1 = \frac{MS_{treatment}}{MS_{error}} = \frac{N(\overline{Y}_E - \overline{Y}_C)^2}{4\hat{\sigma}^2}\,, \tag{9}$$

where $\hat{\sigma}^2$ is the usual pooled variance estimator. Under model (7), F_1 has an *F*-distribution with 1 and *N*-2 degrees of freedom and noncentrality parameter $\lambda_1 = \frac{\beta_1^2}{Var(\hat{\beta}_1)} = \frac{N\beta_1^2}{4\sigma^2}$, which we denote as $F_1 \sim F(1, N\text{-}2, \lambda_1)$. The parameter λ_1 may be written as a function of the effect size $\delta = \beta / \sigma$:

$$\lambda_1 = N\delta^2 / 4. \tag{10}$$

The power of this test is $\Pr[F_1 > F_{1-\alpha}(1, N-2) \,|\, H_a],$ where $F_{1-\alpha}(1, N-2)$ is the 1 - α percentage point of the central-*F* distribution.

Analysis Using Perfectly Measured Fidelity

Now suppose that fidelity varies from one subject to another in the experimental group and is added as a covariate in the analysis. If fidelity were measured perfectly (i.e., κ_k = 1) and was related to outcome, then the power of the test of hypotheses (8) could be increased by fitting the model

$$Y_i = \beta_0 + \beta_1 W_i + \gamma f_i + r_i . \tag{11}$$

To ease interpretation here and in further discussions in this paper, we assume that f_i has been centered around its treatment group mean. That is, fidelity is centered separately within treatment and control groups, if it is measured in the control group. If it is not, then f_i is defined to be 0 for persons in the control group. Even if fidelity is measured in the control group, it should vary little around 0, if the measure is well selected and operationalized to capture the unique aspects of the intervention.

The residual r_i has smaller variance in model (11) than in (7) if outcome is related to fidelity. Using the notation of Spybrook et al. (2011, p. 24), we denote the distribution of the residual for those receiving the intervention as $r_i \sim N(0, \sigma^2_{|f}) = N(0, (1-\rho^2_{fy})\sigma^2),$ where ρ_{fy} is the correlation between fidelity and outcome. However, in calculation of power, we assume that fidelity is nearly constant (at the centered value of 0) in the N/2 cases in the control group, and thus the average residual variance is reduced by only half of what it would be if the covariate reduced variance equally in both groups: $\sigma^2_{|f} = \sigma^2(1-\rho^2_{fy}/2)$. This is a conservative approach from the point of view of power, since this model for $\sigma^2_{|f}$ will underestimate reduction in variance if fidelity does vary in the control group. Often in educational research, overlap between the treatment and the control condition is expected.

The usual F-statistic for testing hypotheses (8) under model (11) has distribution $F_{1|f} \sim F(1, N-3, \lambda_{1|f}),$ where the noncentrality parameter

$$\lambda_{1|f} = N\delta^2 / 4(1-\rho_{fy}^2/2). \tag{12}$$

The power of this test is

$$\Pr[F_{1|f} > F_{1-\alpha}(1, N-3) \mid H_a]\,. \tag{13}$$

Besides testing for a treatment effect, the analyst may also want to test whether or not there is a relationship between fidelity and outcome. Support for the theory on which the intervention is based will be stronger when outcome and fidelity measures of its important components are positively correlated. This evidence will be especially valuable when treatment effect size is small. A test for a relationship between fidelity and outcome; i.e. of

$$H_0 : \gamma = 0 \text{ against } H_a : \gamma \neq 0\,, \tag{14}$$

is based on the statistic $F_{1\gamma} = \hat{\gamma}^2 / [\hat{\sigma}_{|f}^2 / \sum_i (f_i - \bar{f})^2] = \hat{\gamma}^2 / [\hat{\sigma}_{|f}^2 / \sum_{i=1}^{N/2} f_i^2]$, where $\hat{\gamma}$ and $\hat{\sigma}_{|f}^2$ are the usual estimators of the regression coefficient and residual variance from model (11). The last equality follows because f_i is centered, and since we are (conservatively) assuming that fidelity is near 0 for all persons in the control group. Then $F_{1\gamma} \sim F(1, N-3, \lambda_{1\gamma})$, with

$$\lambda_{1\gamma} = (N/2)\rho_{fy}^2 / (1-\rho_{fy}^2)\,. \tag{15}$$

The power of this test can be approximated by

$$\Pr[F_{1\gamma} > F_{1-\alpha}(1, N-3) \mid H_a]\,. \tag{16}$$

Analysis Using Estimated Fidelity

If the reliability of the fidelity measure is not perfect (i.e., $\kappa_k < 1$), then the power functions shown in (13) and (7) are too optimistic. This will occur when only a subset of sessions is observed for fidelity. In that case, if the analyst wants to use the fidelity measure as a covariate, its estimated value will have to replace its unknown true value in model (11). Then the model becomes:

$$Y_i = \beta_0' + \beta_1 W_i + \gamma' \ \hat{f}_i + r_i' \ . \qquad (17)$$

As noted in Section 2, the least-squares estimator of β_1 fit from model (17) is still unbiased, since $\hat{f}_i$ is orthogonal to the treatment indicator (due to separate centering in each group). However the estimator of the coefficient of $\hat{f}_i$ has expectation $\gamma' = \kappa_k \gamma$ (Fuller (1987), eqn. (1.1.7)), where κ_k is the reliability of the fidelity measure as defined in (6). Thus $\hat{\gamma}'$ is biased for γ if reliability is less than perfect. In addition, the correlation between the covariate and outcome is reduced; specifically (from Fuller (1987), eqn. (1.1.17)),

$$\rho^2_{\hat{f}y} = \kappa_k \rho^2_{fy} \ . \qquad (18)$$

As a result, the average variance of r_i' is increased (if $\kappa_k < 1$) over that of model (11) to $\sigma^2_{|\hat{f}} = \sigma^2 (1 - \kappa_k \rho^2_{fy} / 2)$. Despite this, the tests for the hypotheses shown in (8) and (14) are still valid when $\hat{f}_i$ follows the CME model (4). That is, they retain their nominal significance level α, since under *H₀*, because in that case, $\gamma' = \gamma$ and $\sigma^2_{|\hat{f}} = \sigma^2_{|f}$. However, the power for both of the tests is reduced when *H₀* is not true, as shown below.

Denote the usual *F*-statistic for testing the treatment effect from model (17) by $F_{1|\hat{f}}$. When fidelity follows the CME model (4), $F_{1|\hat{f}} \sim F(1, N-3, \lambda_{1|\hat{f}})$, where

$$\lambda_{1|\hat{f}} = N\delta^2 / 4(1 - \kappa_k \rho^2_{fy} / 2). \qquad (19)$$

The noncentrality parameter is smaller because the fidelity measure is now not so strongly correlated with outcome due to its imperfect reliability, as shown in (18). Thus the power for this test is reduced to

$$\Pr[F_{1|\hat{f}} > F_{1-\alpha}(1, N-3) \mid H_a]. \tag{20}$$

Similarly, the power for the test of the relationship of fidelity to outcome is reduced as well. Its test statistic $F_{1\gamma'} \sim F(1, N-3, \lambda_{1\gamma'})$ also has its noncentrality parameter reduced to

$$\lambda_{1\gamma'} = (N/2)\kappa_k \rho_{fy}^2 /(1-\kappa_k \rho_{fy}^2), \tag{21}$$

yielding power of

$$\Pr[F_{1\gamma'} > F_{1-\alpha}(1, N-3) \mid H_a]. \tag{22}$$

The power for both tests are functions of reliability of fidelity measurement, κ_k , which in turn is a function of fidelity sample size. Therefore, the power of one or both tests can be used to guide selection of *k*.

Example

In this example, we show how to use the expressions for power in planning the number of sessions that must be sampled for assessing fidelity. First note that the power of the test of a treatment effect from model (17) is a function of the treatment effect size δ, the significance level of the test α, the total number of individuals in the trial *N*, the correlation between fidelity and outcome ρ_{fy} , the intra-cluster (i.e., individual) correlation of fidelity ICC_f, the number of sessions of the intervention each person receives *K* during the course of the intervention, and number of those sessions sampled *k*. The power of the test of a relationship between outcome and fidelity is a function of ρ_{fy} , α, *N*, ICC_f, *K*, and *k*.

A reasonable approach for the user in determining *k* is to examine power as sample size varies from its minimum of 2 to its maximum of *K*, the total number of sessions. Then choose a value of *k* that provides adequate power, given the other study parameters. In most cases, the power curves increase rapidly with sample size and then level off far below their maximum.

Since ICC_f is likely to be least familiar to the analyst, a range of values of ICC_f should be examined.

Consider the following example, based on one described in Section 4.9 of Spybrook et al. (2011). Researchers were planning a study to investigate the effect on achievement of assignment to a new charter school. The district used a lottery to make the assignment, as there was not sufficient capacity to accommodate all interested students. The performance of students was to be evaluated using the Iowa Test of Basic Skills (ITBS), and an effect size of $\delta = 0.25$ for the charter school treatment was considered important to detect. A power analysis for model (7) using Optimal Design (OD) software (Raudenbush et al., 2011) showed that a sample of $N = 504$ (252 in each group) students would provide power of 0.80 to detect this effect size at significance level $\alpha = 0.05$.

Now suppose that the researchers identify one or more components of the delivered instructional methods of the charter school that they believe may be the drivers of the expected improvement in achievement. They expect, however, that implementation will vary somewhat within the school, since the instructional delivery will not be in a controlled setting. They are interested in testing whether or not the components they have identified are indeed associated with the outcome. They plan to select a sample of k of the K=180 instructional hours for each student and measure the fidelity of those components. They will include each measure as a covariate in the analysis of treatment effect for ITBS score by fitting model (17). They would like to be able to detect an effect size for each relationship if it is at least $\rho_{fy} = 0.20$. How large must k be?

Figure 1 shows the power of this test as a function of k computed as in (22), where $\alpha = 0.05$, $N = 504$, $K = 180$, $\rho_{fy} = 0.20$. Because we are not sure about the value of ICC_f, three values are examined: ICC_f = 0.15, 0.30, and 0.45. The figure shows that power increases rapidly with the number of sessions sampled and levels off for rather small sample sizes. The power exceeds 0.70 for k = 8 and exceeds 0.80 for k = 16 for all three values of ICC_f considered.

Next we examine how dependent the needed sample size is on the number of sessions delivered in all. Suppose the intervention sessions are delivered at a different interval other than daily, so that the total number of sessions is either larger or smaller than 180. For example, if the innovative instruction available in the charter school occurred weekly, rather than daily, then K = 36. At the other extreme, suppose there were 4, rather than one, instructional sessions per day, so that K = 4 * 180 = 720. How does this change the required k?

Table 1 shows the value of k needed to achieve a power of 0.80 for K = 36, 180, and 720 total sessions. The table shows that the number of sessions needed is somewhat reduced when K is small, but not proportionately so. The effect is most sensitive to the size of K when ICC_f is small, that is, when fidelity varies much more between sessions than between individuals. The reason that sample size needed is only weakly related to the number of total sessions K is because reliability κ_k, and thus power, is related to K only through the finite population correction factor (1- k/K), as shown in (6). As a rule of thumb, if sampling rate k/K is less than about 5%, an increase in the total number of sessions K has little effect on power. Note specifically that this means that selecting the fidelity sample size to be a constant proportion of the number of sessions would lead to a waste of resources, from the point of view of power. Rather, unless the sampling rate is very small, the absolute magnitude of the number of fidelity sessions sampled determines its effect on power.

If the analysis were to show a relationship between fidelity and outcome, the outcome data could be analyzed for a treatment effect using model (17). Suppose the researchers, based on the previous analysis, decided on a sample size of k = 8 session fidelity measurements in their study. They would like to determine how much the power to detect a treatment effect size of $\delta =$ 0.25 would increase from using their fidelity measure as a covariate, and how much the power is affected by their choice of k. Figure 2 displays the power of a test of the treatment effect (from (19) and (20)) as a function of the correlation between fidelity and outcome ρ_{fy} for the chosen sample size of k = 8 and ICC_f = 0.30. It also displays the power functions for two other sample size choices: k = 2 (the minimum possible sample size) and $k = K$ = 180, which provides perfect reliability in measurement of fidelity. Figure 2 shows that k = 8 provides most of the increase in power that would be available from using fidelity as a covariate.

SAS code for calculating power for these hypothesis tests is provided in the supplementary material. Power for a test of the treatment effect for various values of k can also be computed using OD Software (Raudenbush et al., 2011), with some pre-processing. Specifically, choose Design ⇒Person Randomized Trial ⇒ Single level trial ⇒ Power (or MDES) vs. explained variation by covariate (R2). Then to observe the power for a specific value of k, observe the power for an x-axis value of $\kappa_k \rho_{fy}^2 / 2$, where κ_k is computed from (7). Power for a test of a relationship between fidelity and outcome cannot be calculated using OD software.

Relationship between Fidelity Measurement and Power for Two-Level Cluster Randomized Trials

When groups of individuals, such as those in classrooms or training centers, are randomly assigned together to treatment or control, the proper analysis of outcomes requires a hierarchical model. The groups are referred to here as clusters. We assume that fidelity is measured at the cluster level. For example, fidelity may be a function of the teacher’s behaviors in an instructional session, and therefore impact all the individuals in her classroom. In this section, we address the same question as previously: If the delivery of the intervention takes place in *K* sessions to the clusters of students, how many of those sessions should be sampled for fidelity monitoring?

Analysis Using Perfectly Measured Fidelity

We first outline a model for outcome using fidelity as a covariate, where fidelity is measured with perfect reliability. Next we examine how the model changes when the only available measure of fidelity for each cluster is one estimated from a sample of sessions. Then we show how to examine the relationship between power of relevant hypothesis tests and this sample size. Note that the calculations for power assume a balanced experiment, having the same number of clusters in the experimental and control groups and the same number of students in each cluster.

Let the outcome of person *i* in cluster *j* of a two-level trial be denoted by Y_{ij}. The generalization of model (11) is written in two parts. The level 1, or person-level model is

$$Y_{ij} = \beta_{0j} + r_{ij} , \tag{23}$$

for $j = 1,\ldots,J$ and $i = 1,\ldots,n$, where *n* is the number of people in each of the *J* clusters, and $r_{ij} \sim N(0,\sigma^2)$. The level 2 model for the cluster mean is:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + \gamma f_j + u_{0j} \tag{24}$$

where W_j = ½ for $j = 1,\ldots, J/2$ (experimental clusters) and = -½ for $j = J/2 + 1,\ldots,J$ (control clusters), f_j is fidelity for those in cluster *j*, and $u_{0j} \sim N(0, \tau_{|f})$. (Assume that f_j is centered within each treatment group.) The residual variance for the level 2 model for the experimental clusters is $\tau_{|f} = (1-\rho^2_{f\beta_0})\tau$, where τ is the residual variance for a reduced level 2 model not including the covariate f_j and $\rho_{f\beta_0}$ is the correlation between fidelity and the cluster mean. As in the single level model, we make the conservative assumption that this reduction in variance occurs only for the clusters receiving the intervention, since the fidelity in the control group should vary little. This provides an average residual variance of $\tau_{|f} = (1-\rho^2_{f\beta_0}/2)\tau$ across all clusters.

The hypotheses to test for a treatment effect are

$$H_0 : \gamma_{01} = 0 \text{ against } H_a : \gamma_{01} \neq 0\,. \tag{25}$$

The F-statistic to test (25) has distribution $F_{2|f} \sim F(1, N-3, \lambda_{2|f})$. Its noncentrality parameter is

$$\lambda_{2|f} = \frac{J\gamma^2_{01}}{4\left[\tau_{|f} + \sigma^2/n\right]} = \frac{J\delta^2}{4\left[(1-\rho^2_{f\beta_0}/2)ICC_y + (1-ICC_y)/n\right]}, \tag{26}$$

where $ICC_y = \tau/(\tau+\sigma^2)$ is the intra-cluster correlation of outcome Y_{ij}[1] and $\delta = \frac{\gamma_{01}}{\sqrt{\tau+\sigma^2}}$ is the standardized effect size. The power of this test is $\Pr[F_{2|f} > F_{1-\alpha}(1, N-3) \mid H_a]$.

The hypotheses to test for a relationship between fidelity and mean outcome is the same as that shown in (14). The F-statistic for testing this hypothesis has distribution $F_{2\gamma} \sim F(1, N-3, \lambda_{2\gamma})$, with noncentrality parameter

$$\lambda_{2\gamma} = \frac{J}{2}\frac{\gamma^2\sigma^2_f}{\left[\tau_{|f} + \sigma^2/n\right]} = \frac{J}{2}\frac{\rho^2_{f\beta_0} ICC_y}{\left[(1-\rho^2_{f\beta_0})ICC_y + (1-ICC_y)/n\right]}. \tag{27}$$

Its power is $\Pr[F_{2\gamma} > F_{1-\alpha}(1, N-3) \mid H_a]$.

Analysis Using Estimated Fidelity

We now consider the case when f_j is not observed directly, but is estimated for each cluster from a sample of k of the K instructional sessions. The estimator $\hat{f}_j$ is assumed to be

[1] It is important to distinguish ICC_y, which is the proportion of variation across all person outcomes that is due to the cluster and ICC_f which is the proportion of variation across all instructional sessions that is due to cluster. The total variance is being measured over different units in the two cases.

related to the true fidelity f_j by the CME model, as defined in (4), and is used as the predictor in the level 2 model:

$$\beta_{0j} = \gamma'_{00} + \gamma_{01}W_j + \gamma'\hat{f}_j + u'_{0j}\ . \tag{28}$$

As in the single level model, the regression coefficient for $\hat{f}_j$ is biased for γ , with expectation

$$E(\hat{\gamma}') = \kappa_k \gamma, \tag{29}$$

where κ_k (defined in (6)) is the reliability of $\hat{f}_j$. The usual least squares estimator of γ_{01} is still unbiased despite the measurement error since $\hat{f}_j$ is defined (due to centering in both treatment groups) to be independent of W_j. However, the average residual variance of $u_{0\text{j}}$ is larger when reliability of fidelity is not perfect, compared with when κ_k .= 1.

As a result, the F-tests for the significance of γ_{01} and γ are calculated in the same way as before, but substituting $\hat{f}_j$ for f_j in the model. The resulting F-statistics, denoted by $F_{2|\hat{f}}$ and $F_{2\gamma'}$, both retain their significance level. However they lose power compared to when κ_k = 1, since the amount of variation in β_{0j} explained by $\hat{f}_j$ is less than the amount explained by f_j . Specifically, $\rho^2_{\hat{f}\beta_0} = \kappa_k \rho^2_{f\beta_0}$ and the residual variance in the level 2 model is $\tau_{|\hat{f}} = (1-\kappa_k \rho^2_{f\beta_0})\tau$. Thus the power for the test of the treatment effect is

$$\Pr[F_{2|\hat{f}} > F_{1-\alpha}(1, J-3,) \,|\, H_a], \tag{30}$$

where $F_{2|\hat{f}} \sim F(1, J-3, \lambda_{2|\hat{f}}),$ and

$$\lambda_{2|\hat{f}} = \frac{J\gamma_{01}^2}{4\left((1-\kappa_k \rho^2_{f\beta_0}/2)\tau + \sigma^2/n\right)} = \frac{J\delta^2}{4\left((1-\kappa_k \rho^2_{f\beta_0}/2)ICC_y + (1-ICC_y)/n\right)}\ . \tag{31}$$

The power for the test of a relationship between fidelity and outcome is

$$\Pr[F_{2\gamma'} > F_{1-\alpha}(1, N-3) \mid H_a], \qquad (32)$$

where $F_{2\gamma'} \sim F(1, N-3, \lambda_{2\gamma'}),$ and

$$\lambda_{2\gamma'} = \frac{J}{2}\left[\frac{\gamma^2\sigma_f^2}{\tau_{|\hat{f}} + \sigma^2/n}\right] = \frac{J}{2}\left[\frac{\kappa_k\rho_{f\beta_0}^2 ICC_y}{(1-\kappa_k\rho_{f\beta_0}^2)ICC_y + (1-ICC_y)/n}\right]. \qquad (33)$$

The power of each test is a function of *k*, so can be used for guiding the sample design.

Example

We illustrate with an example the process of considering the sample size needed for fidelity measurement from the point of view of power of tests of the outcome variable. The example is based on one described in Section 7.9 of Spybrook, et al. (2011). Suppose researchers are preparing to investigate the effect on achievement of a new literacy program for third graders. They will recruit schools to participate, and assign half to the new program and half to the control program. The literature suggests that 20% of the variability in student reading achievement is between schools (ICC_y = 0.20). They plan to assign one classroom of 20 students in each school to the study (*n* = 20). They are interested in detecting an effect size δ of 0.25. OD software was used to conduct a power analysis, and showed that a sample of *J* = 122 schools, 61 in each group, would be needed to provide power of 0.80 to detect this effect using a test with significance level $\alpha = 0.05$.

Now suppose that researchers identify one or more components of the delivered instructional methods of the reading program that they believe may be the drivers of the expected improvement in achievement. They expect, however, that implementation will vary somewhat between schools. They are interested in testing whether or not the components they have identified are associated with outcome. They plan to select a sample of *k* of the *K*=180 instructional hours for each participating classroom and measure the fidelity to those components and to include each measure individually as a covariate in the analysis of treatment effect for reading achievement by fitting model (23) - (24). They want to detect a relationship if its effect size $\rho_{f\beta_0}$ is at least moderate, which they consider to be 0.35. They are also interested, of course, in detecting the treatment effect, so would like to consider how much it can increase with the use of fidelity in the analysis, if the correlation is indeed of magnitude 0.35. How large should *k* be to accomplish these goals?

Figure 3 shows the power functions for both the test of a treatment effect and a fidelity relationship as functions of *k*, where $\alpha = .05$, $J = 122$, $K = 180$, $n = 20$, $\rho_{f\beta_0} = 0.35$, $ICCy = 0.20$, and $ICC_f = 0.15$ and 0.45, which represent two quite different levels of consistency of fidelity delivery. The figure shows that the increase in power for detecting the treatment effect is negligible, no matter the size of *k*. This occurs because only 20% of the variability in outcome is due to variability among clusters, the unit of measurement for fidelity. Thus there is not much to be gained by reducing this minor portion of variability by using fidelity as a covariate. However, the figure also shows that power of the test of the relationship between fidelity and outcome is greatly affected by *k*. It increases rapidly and then levels off, for both values of ICC_f.

When $ICC_f = 0.45$, a small sample of about $k = 10$ of the 180 sessions are sufficient to realize nearly all the power available. When $ICC_f = 0.15$, a larger sample is needed. When ICC_f is large, it means that each school has a similar level of fidelity across sessions. When ICC_f is small, it means that all schools deliver similar average fidelity over the entire study period, though it may vary from session to session within the school. It is obvious that if a single school delivers the treatment with near identical fidelity over all sessions, then a small number of fidelity measurements will provide high reliability in the fidelity measure. This illustrates the importance of knowing the consistency of fidelity across sessions for optimal design of data collection. If the fidelity measure is new, ICC_f is unlikely to be known. However, it can be measured during the experiment itself, which can help in future studies.

SAS code for calculating power for this example is provided in the supplementary material. Power for a test of the treatment effect can also be computed using OD Software (Raudenbush, et al., 2011), with some pre-processing. Specifically, choose Design ⇒Cluster Randomized Trial with person-level outcome ⇒Cluster randomized trials ⇒Treatment at level 2 ⇒ Power (or MDES) vs. proportion of variation explained by level 2 covariate (R2). Then to observe the power for a specific value of k, observe the power for an x-axis value of $\kappa_k \rho^2_{f\beta_0} / 2$, where κ_k is computed from (7). Power for a test of a relationship between fidelity and outcome cannot be calculated using OD software.

Discussion and Implications for Practice

In this section, we address some of the practical issues that the researcher will encounter when carrying out the proposed power analysis and implementing the fidelity sample design chosen. We also make recommendations for improving how fidelity measures are collected and used.

Selection of Parameter Settings

In order to use the method outlined in this paper for designing the fidelity sample, assumed values for a variety of parameters are required. Most of these are familiar, since they are also required for planning for a power analysis of the treatment effect itself. For example, researchers are by now accustomed to making choices about effect size for the treatment effect, due to guidance to social scientists initially provided by Cohen (1992). He also provided guidance about the expected effect size for covariates in a regression. Hedges and Hedberg (2007) provided guidance on how to choose ICC_y when designing experiments with hierarchical structure in educational research. Their survey of the literature showed that most such studies have reported values in the range of 0.15 to 0.25 for educational outcomes, where clusters are schools.

The required parameter that is least familiar to most researchers is ICC_f. To our knowledge, there have been no reports in the literature of ICC_f values for fidelity measures. Since measures of fidelity are so diverse, it may not be possible to arrive at a range, like Hedges and Hedberg did, that would be appropriate for all. However, if researchers did begin reporting these values, it may be possible to draw conclusions about certain types of common fidelity measures.

The parameter ICC_f reflects the proportion of the total variability in fidelity that is between clusters (in a cluster randomized design) or between students (if in a person randomized design). That is, it is the analog of *ICCy*, which describes the same concept for outcomes. ICC_f can be estimated using software fitting a one-way ANOVA with random effects, with fidelity as the response variable and cluster or student as the class variable.

If the researcher has no fidelity data available from pilot studies, he or she may still use the method outlined to provide a range of values for *k*. For example, he or she many select low (say $ICC_f = 0.15$) and high (say $ICC_f = 0.50$) values for ICC_f and calculate the range of values required for *k* under these conditions.

Selection of Fidelity Sample

The variance of any estimator depends on the sample design. The variance expression for $\hat{f}_i$ in (3) is based on estimation from a random sample of intervention sessions. In some situations, this may be a feasible sample design. For example, in an experiment in which all intervention sessions were videotaped, but only a subset coded for fidelity due to resources, the selection of the sessions to be coded can easily be chosen randomly. In other circumstances, such as when observation for fidelity must be conducted in real time, random selection of sessions would likely not be practical, due to a need for predictable staffing and a desire to ensure coverage over the entire period of the intervention.

One approach for handling this problem is to use a systematic random sample design for fidelity monitoring. To implement this design for a sample of size *k*, first select a session at random from the interval between 1 and the integer nearest *K/k*, denoted by *r*. Then select that session and every r^{th} session thereafter into the sample. This design has the advantage of evening the workload over the period of the intervention, and can also provide a better estimator (i.e., smaller variance) of mean fidelity if fidelity varies smoothly over the intervention period. A disadvantage of this method is that an unbiased estimator of the variance of the fidelity measure is not available without making some assumptions about how fidelity varies over time. The expression in (3) will overestimate the variance if the fidelity improves or gets worse steadily over the course of the intervention. However, it probably is best to overestimate rather than underestimate the variance as it provides a conservative p-value for both hypothesis tests.

Subsampling of sessions

Another practical question about the sample design is how to define the fidelity sampling unit. The unit has been referred to as a “session” in this paper. However, researchers may define an observation session in any way they choose; e.g., as a quarter-hour rather than the hour session in which the intervention is delivered. In fact, defining shorter fidelity observation periods may result in a more efficient design, if consistency of fidelity within the intervention session is high. The power calculation only requires adjustment of the total number of sessions, *K*, as well as revisiting assessment of ICC_f. Note that the sample of *k* sessions, however they are defined, must still be selected randomly.

In summary, this article addresses the practical issue of determining an adequate sample size for fidelity observations conducted as part of randomized control trials. Just as researchers conduct power analyses to determine the number of participants needed for their trials, we provide procedures for determining the number of observations needed to adequately estimate fidelity and we provide a mathematical rationale to support these procedures. The practical need for this work is clear as conducting fidelity observations requires extensive resources (Gersten, et al., 2000). As researchers strive to more adequately assess fidelity of implementation, using the proposed power analysis will provide researchers with a reasoned approach for determining the number of observations needed.

Recommendation for future development

The analysis outlined in this paper can be implemented in a fairly straightforward way using any statistical software containing a non-central *F* function, such as SAS, SPSS, or STATA. The more perplexing problem is how to predict the settings for the parameters required as inputs to the analysis. Cohen (1992) may have been the first to try to provide guidance to researchers by his specification of small, medium, and large effect sizes for various types of parameters. Recently, there have been several useful surveys of the empirical literature from educational research studies to determine what reasonable values of ICC_y, could be expected in educational studies. Examples of these are Hedges and Heberg (2007) and Westine, Spybrook, and Taylor (2014). In order to improve the design of fidelity data collection, planning values for ICC_f are also needed. We recommend that future researchers who carry out careful fidelity measurement in their intervention studies routinely compute and make available estimates of their means, within and between-individual variances, and ICC_f values.

This paper has described an approach to determine the number of fidelity measurements needed for two types of intervention designs, the person randomized and two-level cluster randomized. For each design, only one design for collecting fidelity data was considered. For example, we assumed that in the cluster randomized design, fidelity measurement will be made at the cluster level, and will apply to all subjects within the cluster. There are clearly many other possible designs for both the study and fidelity data collection.

Munter et al. (2014) describe a study in which students were assigned to one of 18 tutors (clusters), but fidelity measurement was made at the student level. Neither of our power analyses would apply to their fidelity sample design because it differed from those we considered. Their design was two-stage; specifically, they randomly sampled about six students (approximately one- third) from each tutor, and then assessed a randomly chosen sample of 12 sessions from each. Besides the question of sample size itself, a relevant question for this design is how the sample could best be allocated. That is, would the approximately 18*6*12 = 1296 sampled and coded sessions have produced higher power for the tests of the treatment and/or implementation effect if they had been allocated across the three levels in a different way? The topic of optimal design for measurement of outcomes is well studied (e.g., see Raudenbush (1997) and Raudenbush and Liu (2000)), but has not been addressed for fidelity data collection. In addition, Munter et al. (2014) also documented that the uncertainty in their measurements of fidelity were due not only to sampling, but also due to imperfect reliability of the fidelity measurements made by coders. Guidance on reliability requirements could also be addressed by using the objective of adequate power as a guide. This example illustrates that more work is needed in how to design of fidelity data collection.

Table 1

Required k to Achieve a Power of 0.80 for Test of Relationship Between Fidelity and Outcome Under Model (16)[a]

	K		
ICC_f	36	180	720
0.15	k = 12	k = 16	k = 17
0.30	k = 6	k = 7	k = 7
0.45	k = 4	k = 4	k = 4

[a] for effect size ρ_{fy} = 0.20, for several sizes of K and ICC_f

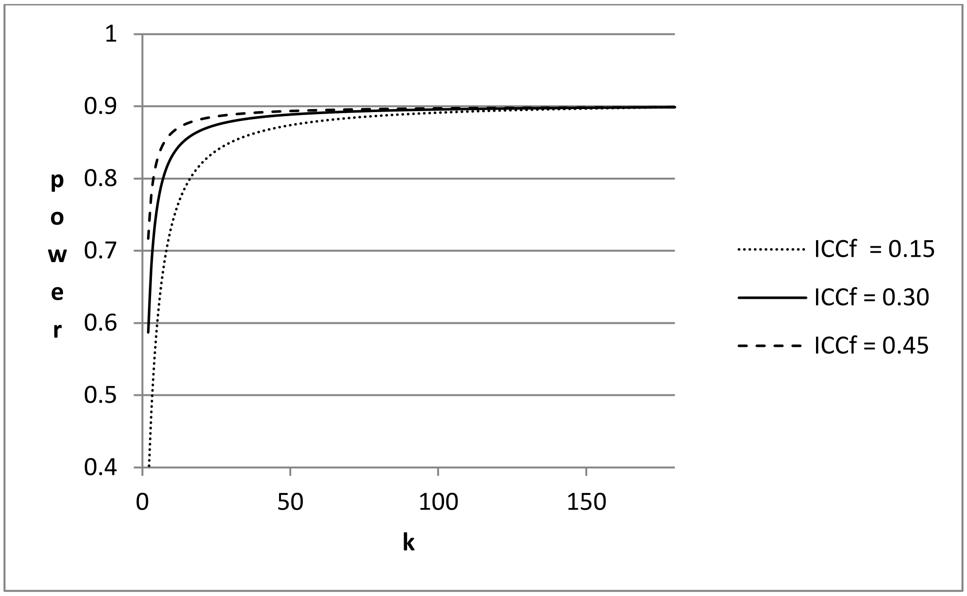

Figure 1. Power for test of outcome's relationship with fidelity as a function of fidelity sample size k, in a person-randomized trial, for ρ_{fy} = .2, K = 180, α = .05, N = 504, and 3 values of ICC_f.

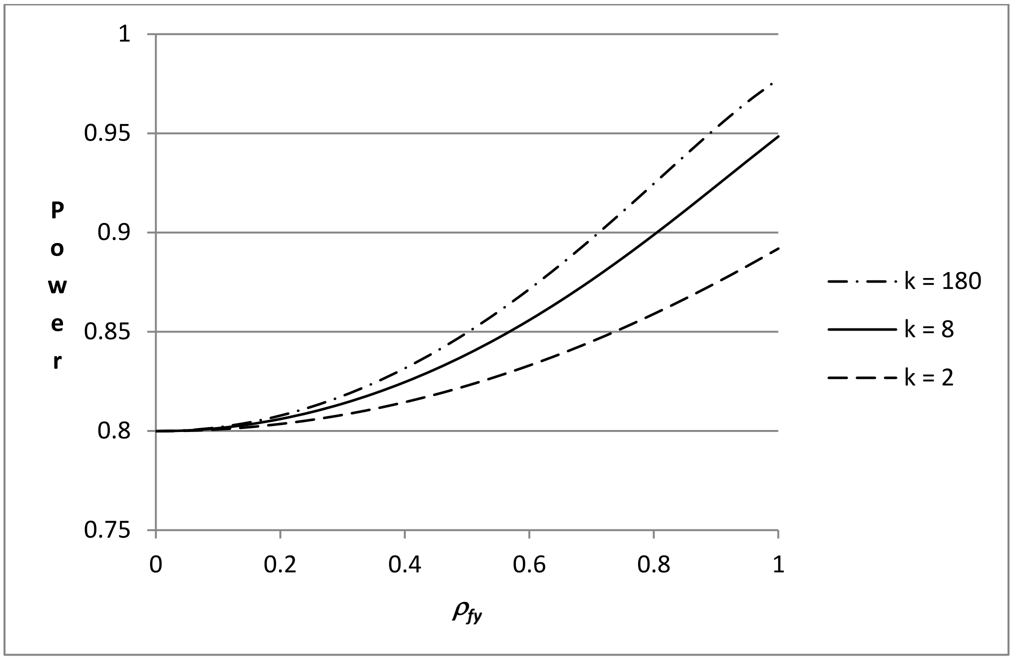

Figure 2 Power for test of treatment effect in a person-randomized trial as a function of correlation between fidelity and outcome ρ_{fy}, for δ = .25, *K*= 180, α = 0.05, *N* = 504, ICC_f = 0.30, and 3 values of *k*.

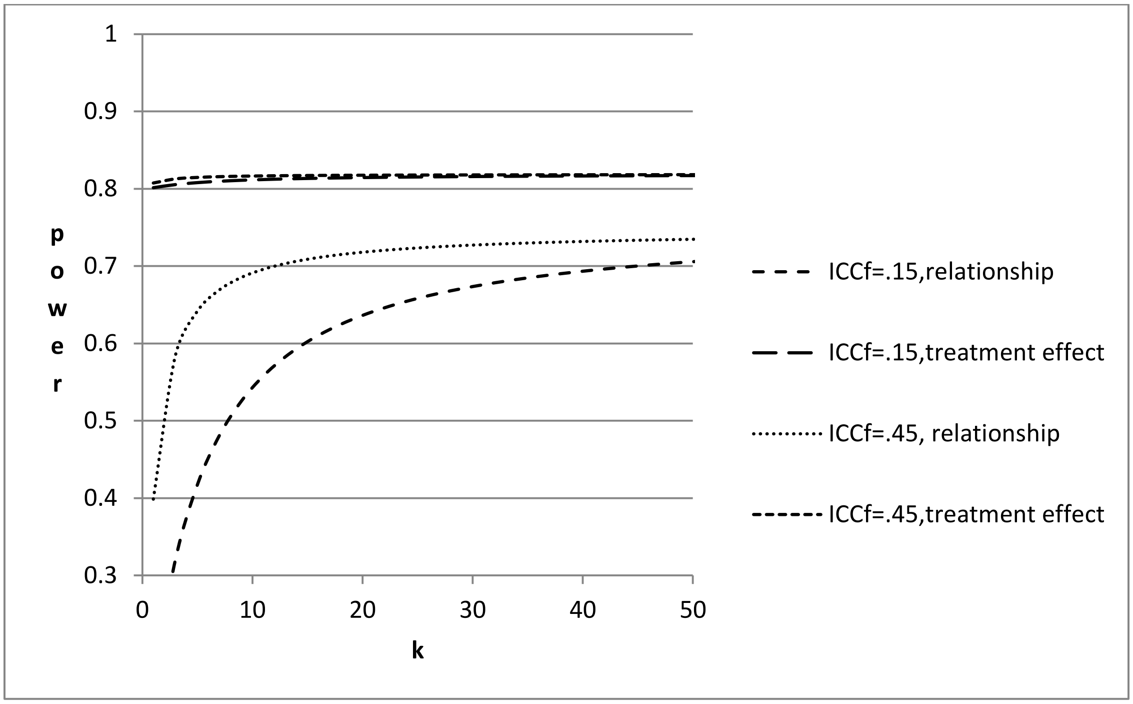

Figure 3. Power for tests of relationship between fidelity and outcome and treatment effect for cluster randomized design as functions of *k*, for *J* = 122, n = 20, *K* = 180, α = .05, δ= .25, and 2 values of ICC_f.

References

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

Cohen, J. (1992). A Power Primer, *Psychological Bulletin*, 112, 155-159.

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

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3-4), 327-50. doi:10.1007/s10464-008-9165

Fuller, W. A. (1987) *Measurement error models*. New York: John Wiley and Sons.

Gall, M. 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.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*(2), 149-164.

Hedges, L. and Hedberg, E.C. (2007). Intraclass correlation values for planning group randomized trials in education, *Education Evaluation and Policy Analysis* 29: 60-87.

Hill, A. B. (1953). Observation and experiment. *New England Journal of Medicine*, *2*48(24), 995-1001.

Lohr, S. L. (2010). *Sampling : Design and analysis* (2nd ed.). Boston, Mass.: Brooks/Cole

Munter, C., Wilhelm, A. G., Cobb, P., & Cordray, D. S. (2014). Assessing fidelity of implementation of an unprescribed, diagnostic mathematics intervention. *Journal of Research on Educational Effectiveness*, *7*(1), 83-113.

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services & Research*, *39*(4), 374-96. doi:10.1007/s11414-012-9295-x

ODonnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K--12 curriculum intervention research. *Review of Educational Research*, *78*(1), 33-84.

Raudenbush, S.W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2), 173-185.

Raudenbush, S.W. and Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5(2), 199-213.

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X. -F., Martinez, A., & Bloom, H. (2011). Optimal design software for multi-level and longitudinal research (version 3.01)[software]. *Available From www.wtgrantfoundation.org*

Spybrook, J., et al. (2011). Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software Version 3.0. Available from www.wtgrantfoundation.org.

Smith, S. W. & Daunic, A. P. & Taylor, G. G.(2007). Treatment Fidelity in Applied Educational Research: Expanding the Adoption and Application of Measures to Ensure Evidence-

Based Practice. *Education and Treatment of Children* 30(4), 121-134.

Westine, C. D., Spybrook, J., & Taylor, J. A. (2014). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review, 37*(6), 490-519. doi:10.1177/0193841X1453158